

Copyright
by
Mehul Pravin Sampat
2006

The Dissertation Committee for Mehul Pravin Sampat
certifies that this is the approved version of the following dissertation:

**Evidence-based Detection of Spiculated Lesions on
Mammography**

Committee:

Alan C. Bovik, Supervisor

Mia K. Markey, Supervisor

John A. Pearce

John E. Gilbert

Gary J. Whitman

**Evidence-based Detection of Spiculated Lesions on
Mammography**

by

Mehul Pravin Sampat, M.S, B.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2006

Dedicated to

My parents Madhavi and Pravin Sampat

My wife Pallavi

My sister Ameer

Acknowledgments

I would like to thank my parents for their constant support and encouragement. Without their sacrifices I would not have been here and this dissertation could not have been completed. My wife Pallavi has always been very supportive of my decision to pursue a doctoral degree and she is a constant source motivation for me!

I would like to thank a number of people without whose support and guidance this work could not have matured to its current state. First of all I want to thank my advisor Prof. Al Bovik who has been a wonderful mentor. His direction and guidance have been invaluable to me and have helped shape me as a researcher. His guidance has helped me develop a scientific and principled approach to problem solving. I would like to thank him for his patience with me and for his belief in me.

I would also like to thank my co-advisor Prof. Mia Markey for her support and guidance. I am here first Ph.D. student and it has been a great pleasure to work with her. Her guidance has been invaluable to me and has made a great impact on me. Mia has also helped me develop my mentoring skills by allowing me to co-supervise various undergraduate students at BMIL. These students have worked on a number of different projects and it has been a wonderful to see them succeed!.

I would like to thank my committee members Dr. Pearce, Dr. Gilbert and Dr. Whitman for their suggestions and criticisms during various stages of my work. Your feed-back has helped to improve the quality of this work.

I have wonderful memories of my time spent at LIVE and I will miss all of the LIVE members a lot. I would like to thank all LIVE members: Umesh, Sumohana, Kalpana, Hamid, Farooq, Hyohoon, Joonsoo, Shalini, Raghu, Yang, James, Abtine and Sina. All of you have made the stay at LIVE very special and I will always cherish these memories. There were so many good memories that I must recount some of them!. I enjoyed playing table-tennis with Farooq even though he currently leads the series (300 to 10). Playing tennis with Kalpana, Sumohana Umesh and Balaji has been a lot of fun though I regret that I due to my laziness i did not go swimming with y'all. Hamid created this stupid song and jingle with my last name and I still don't know why he finds it so funny. As you may read in the acknowledgments of many recent dissertations at LIVE, Umesh deserves a special mention for all of the ways he has selflessly helped LIVE members.

I would also like to thank the members of BMIL for their support. I have enjoyed the company of Ernest, Jin, Min, Qiu, Wendy and Sunyoung. It has also been a lot of fun working with the various under-graduate students, Noah, Amit, Nitin, David, Tony and Yuhling. Your passion for 'pineapple pizza' was very infectious. I would also like to thank Zack and Chris for their technical support.

I would also like to thank the staff members of the BME and ECE

departments. Majid, Cheryl, Joni, Heidi, Ann and Melanie have helped in various administrative tasks. I am sure I have missed someone or the other and I thank all of you who have either directly or indirectly helped me during my stay in Austin.

Evidence-based Detection of Spiculated Lesions on Mammography

Publication No. _____

Mehul Pravin Sampat, Ph.D.
The University of Texas at Austin, 2006

Supervisors: Alan C. Bovik
Mia K. Markey

The American Cancer Society estimates that 212,920 women will be diagnosed with breast cancer in the United States in 2006. Another 40,970 women will die of the disease. Early detection of breast cancer increases the survival rate and increases the treatment options. Screening mammography, radiographic imaging of the breast, is currently the most effective tool for early detection of breast cancer. Radiologists visually search mammograms for specific abnormalities. Some of the important signs of breast cancer that radiologists look for are clusters of micro calcifications, masses, and architectural distortions.

However, mammography is not perfect. Detection of suspicious abnormalities is a repetitive and fatiguing task. For every thousand cases analyzed by a radiologist, only three to four are cancerous and thus an abnormality may be overlooked. Radiologists fail to detect 10% to 30% of cancers and two

thirds of these are evident retrospectively. Thus, computer-aided detection (CAdE) systems have been developed to aid radiologists in detecting mammographic lesions that may indicate the presence of breast cancer. However, it is widely known that these systems are more accurate for the detection of micro-calcifications than spiculated lesions.

In this dissertation a new *evidence-based* algorithm is developed for the detection of spiculated lesions on mammography. By evidence based, we mean that we use the statistics of the physical characteristics of these abnormalities to determine the parameters of the detection algorithm. Towards this goal, we have shown that the properties of these lesions can be measured reliably and we have created the first database of the physical properties of these lesions.

For the detection algorithm, we have invented a new class of linear filters and filter banks which we call Spiculation Filters and Spiculation Filter banks. These filters were created specifically for the detection of spiculated lesions and are highly specific narrowband filters, which are designed to match the expected structures of these abnormalities. As a part of this algorithm, we have also invented a novel technique to enhance spicules on mammograms. This entails filtering in the Radon domain. All the parameters of the detection algorithm are based on measurements of physical properties of spiculated lesions. The results of the detection algorithm are presented in the form of FROC curves and are competitive with existing algorithms.

Table of Contents

Acknowledgments	v
Abstract	viii
List of Tables	xiii
List of Figures	xv
Chapter 1. Introduction	1
1.1 Motivation	1
1.2 Computer-Aided Detection of Mammographic Abnormalities .	4
1.2.1 Detection of Masses	5
1.2.1.1 Stage 1: Detection of Suspicious Regions	6
1.2.1.2 Stage 2: Classification of Suspicious Regions as Mass or Normal Tissue	17
1.2.1.3 Methods for the detection of architectural distor- tions	18
1.2.2 Commercial Computer-Aided Detection Systems	19
1.2.2.1 R2 Technology, Inc.	19
1.2.2.2 Intelligent Systems Software, Inc.	19
1.2.2.3 CADx Medical Systems	20
1.2.3 Independent Studies of Commercial Computer-Aided De- tection Systems	20
1.3 Overview of dissertation	23
Chapter 2. Measurement Studies: The reliability of measur- ing physical characteristics of spiculated masses on mammography	31
2.1 Introduction	31
2.2 Materials and Methods	33

2.2.1	Data Set	33
2.2.2	Observer Training and Measurement Protocol	34
2.2.3	Statistical Analysis	38
2.3	Results	40
2.4	Discussion	44
Chapter 3.	Detection Algorithm	49
3.1	Introduction	49
3.2	Overview of Spiculation Filtering	50
3.2.1	Enhancement of Spiculations	51
3.2.2	Detection Using Spiculation Filters	56
3.3	Detecting central mass regions	64
3.4	Suppression of linear structures	65
3.5	Methods	66
3.5.1	Data Description	66
3.5.2	Selection of Detection Algorithm Parameters Based on Measurement Data	69
3.6	Results	72
3.7	Discussion	73
3.7.1	Comparison with previous studies	73
Chapter 4.	Comparison of Spicule Enhancement Algorithms	78
4.1	Introduction	78
4.2	Materials and Methods	78
4.2.1	Brief review of spicule enhancement algorithm	78
4.2.2	Experimental Setup	80
4.2.3	Results	81
4.3	Discussion	82
Chapter 5.	Similarity Metrics for Binary Images: Comparing Segmentations on Medical Imaging	87
5.1	Introduction	87
5.2	Metrics	95
5.2.1	Intensity-Based Similarity Metrics	95

5.2.2	Spatial-Based Similarity Metrics	98
5.2.2.1	Pixel Correspondence Metric (PCM)	100
5.2.2.2	Closest Distance Metric (CDM)	101
5.2.3	Complex Wavelet Structural Similarity (CW-SSIM) . . .	101
5.3	Data Description and Experimental Design	104
5.4	Results	106
5.4.1	Results on Simulated Data	106
5.4.1.1	Effect of rotation	106
5.4.1.2	Effect of translation	107
5.4.1.3	Effect of width of edges	108
5.4.2	Results on Mammography Data	109
5.5	Discussion	111
Chapter 6.	Conclusion and Future Work	123
6.1	Conclusion	123
6.2	Future Work	124
6.2.1	Relations to Spiculation filter-banks to wavelet theory .	124
6.2.2	Local Radon Transform	125
6.2.3	Measuring properties of other linear structures in mam- mograms	126
6.2.4	Collection of measurement data on spiculated lesions . .	126
Appendices		128
Appendix A.	List of images used for measurement study	129
Appendix B.	List of images used for the detection algorithm	131
Bibliography		135
Vita		154

List of Tables

2.1	Properties of the two sets of images used in this study. A set of 12 cases was used for observer training and measurements were collected using a second, distinct set of 21 cases.	34
2.2	This table provides the summary statistics and the results of the hypothesis test for equivalence between the measurements of radiologists R1 and R2. The null hypothesis was that the two radiologists are not equivalent. Thus, if we obtain a p-value of less than 0.05 (bold type), we can reject the null hypothesis and say that the measurements of the two radiologists are equivalent. (§ One observer measured minor axis by mistake, so that image was removed for the major axis calculation only.)	41
2.3	(Inter-Observer agreement) Intra-class correlation coefficients for the measurements made by radiologist R1 and R2. (§ One observer measured minor axis by mistake, so that image was removed for the major axis calculation only.)	42
2.4	Results of the hypothesis test for equivalence between the first and second set of measurements made by radiologists R1. The null hypothesis was that the two sets of the measurements made by radiologist R1 are not equivalent. Thus, if we obtain a p-value of less than 0.05 (bold type), we can reject the null hypothesis and say that the two sets of measurements are equivalent.	43
2.5	(Intra-Observer agreement) Intra-class correlation coefficients for the two sets of measurements made by radiologist R1 . . .	44
3.1	This table summarizes the performance of prior algorithms developed for the detection of spiculated masses and architectural distortions. It also presents the number of training and testing images used in each study.	74
4.1	This table shows the results of the two-forced alternative choice (2-AFC) experiment. For each observer, the percentage of images where the observer felt the spicules were enhanced more prominently by Method 2 (which used the FSS technique) is shown. On average, for 74% of images the observers felt that Method 2 produced better enhancement results. In comparison, the radiologist preferred the images generated by Method 2, for 28 out of the 30 images (98%).	82

5.1	Two-by-two contingency table: The goal is to compare two segmentations of the same object. A pixel that lies on or inside the boundary of the segmented object is labeled one and zero otherwise. In this table a represents the number of pixels that are non-zero in both images. The term b represents the number of pixels marked as one by segmentation generator #1 only and c represents the number of pixels marked as one by segmentation generator #2 only. Finally, d represents the number of pixels that were labeled as zero by both segmentation generators. . .	96
5.2	Similarity metrics for binary data: These are a set of metrics that are defined based on the terms a , b , and c only (from Table 5.1). The Dice and Jaccard metrics have been commonly used in medical imaging applications whereas the remaining metrics are more popular in other fields. These metrics are all closely related.	98
5.3	Metrics for binary data: These are a set of metrics that are defined based on the terms a , b , c , and d (from Table 5.1). Note that for image segmentation similarity, d represents the number of pixels that were labeled as zero in both segmentations. Thus, these metrics may not be as desirable as those in Table 5.1 since they depend on the size of the background surrounding the segmented objects.	99
A.1	This table lists the set of spiculated masses used in the measurement studies. Each radiologist measured the properties of these spiculated masses.	130
A.2	This table lists the set of architectural distortions used in the measurement studies. Each radiologist measured the properties of these architectural distortions.	130
B.1	This table lists the set of spiculated masses used in the detection studies. The detection algorithm is applied to each of these images and results are reported as FROC curves.	132
B.2	This table lists the set of architectural distortions used in the detection studies. The detection algorithm is applied to each of these images and results are reported as FROC curves.	133
B.3	This table lists the set of spiculated masses (from the MIAS database) used in the detection studies. The detection algorithm is applied to each of these images and results are reported as FROC curves.	134

List of Figures

1.1	In screening mammography two views of each breast are recorded; the craniocaudal (CC) view (left), which is a top-to-bottom view, and a mediolateral oblique (MLO) view (right), which is a side view taken at an angle. The images were obtained from Digital Database for Screening Mammography (DDSM), [62] .	25
1.2	Examples of a spiculated mass (left), cluster of microcalcifications (center), and architectural distortion (right). The images were obtained from the DDSM, [62]	26
1.3	A flowchart showing the main steps involved in the computer-aided detection (CAdE) and computer-aided diagnosis (CAdx) of mammographic abnormalities. Most detection algorithms consist of two stages. In stage 1, the aim is to detect suspicious lesions at a high sensitivity. In stage 2, the aim is to reduce the number of false positives without decreasing the sensitivity drastically. The steps that are involved in designing algorithms for both stages are shown in (b). We note that in some approaches some of the steps may involve very simple methods or be skipped entirely. For example, in stage 1, the classification step often is a simple size criteria (i.e., if the size of potential lesion is suspicious only if its size is greater than N pixels). Most diagnosis algorithms (CAdx) begin with a region of interest (ROI) containing the abnormality. Again, the steps typically involved in design such a system are shown in (b). The output of a CAdx system may be the likelihood of malignancy or a management recommendation. Different research groups have worked on different components of the problem and human interaction may occur at various stages. For example, many CAdx algorithms start with manually segmented ROIs.	27
1.4	Free-response receiver operating characteristic (FROC) curve: In an FROC curve, sensitivity is plotted on the y-axis and the number of FPI is plotted along the x-axis. FROC curves are used to report the performance of detection studies.	28
1.5	(a) Directions of spicules of a spiculated lesion differ from the directions of normal linear markings in a mammogram and the (b) standard deviation of the gradient orientation histogram differentiates the area near a spiculated lesion from normal. (©2006 IEEE).	28

1.6	Three-directional second-order Gaussian derivatives used for estimation of line orientation. The figure was obtained from Prof. Nico Karssemeijer (©2006 IEEE).	29
1.7	Partitioning of the frequency domain achieved with the directional wavelet transform (©2004 IEEE).	30
2.1	NIH ImageJ Interface for obtaining measurements of key characteristics of spiculated masses.	36
2.2	Examples of the measurements made by radiologists R1 and R2 before and after the training stage. The measurements made by R1 are shown in blue and those made by R2 are shown in green.	37
2.3	Bland-Altman analysis for the inter-observer agreement for each of the four physical characteristics that were measured by radiologists R1 and R2. The parameters measured were: (a) major axis of the spiculated masses, (b) the width of the spiculations, (c) the length of the spiculations, and (d) the number of spiculations.	47
2.4	Bland-Altman analysis for the intra-observer agreement for each of the four physical characteristics that were measured twice by the senior radiologist R1. The parameters measured were: (a) major axis of the spiculated masses, (b) the width of the spiculations, (c) the length of the spiculations, and (d) the number of spiculations.	48
3.1	This figures shows the block diagram of the detection algorithm. This algorithm consists of three components. The first component aims to detect the spatial locations where spicules converge. The second component is designed to detect the central mass regions of the spiculated masses and the third component aims to reduce false-positives due to normal linear structures.	50
3.2	These images show lines of different widths and the corresponding section of their Radon transforms. For simplicity, we assume that each of these lines is 255 pixels long and the intensity at each pixel is one. Thus, peaks of value 255 occur in the Radon domain. A key point is that lines of different widths have different representations in the Radon domain.	54
3.3	Flow-chart of the Enhancement Algorithm: First, the Radon transform of an image is computed. The Radon domain is then filtered with a column filter. Finally, the enhanced image is obtained by computing the inverse Radon transform.	56

3.4	Example of a Spiculation Filter (SF): Figures 3.4(a) and 3.4(b) show the two quadrature components of the SF, namely, the cosine SF and the sine SF, respectively. The toroidal Gaussian envelope of the SF is shown in Fig. 3.4(c).	59
3.5	Spiculation FilterBanks (SFBs): Two Spiculation Filter banks, each composed of five SF of progressively increasing radii and matching radial frequencies, ω . The SFB in Figs. 3.5(a) and 3.5(b) have the same inner radius but different outer radii. These would correspond to lesions where the central mass is the same size but the spicule lengths differ.	60
3.6	Figure 3.6(a) shows the toroidal components of two SFs of a filter-bank and Fig. 3.6(b) shows a cross-section through these components. The filters are chosen such that the component toroidal Gaussian envelopes intersect each other at their half-peak values, a strategy that ensures that there will be no gaps in the responses along the radial directions.	62
3.7	This figure shows the set of 4 Gaussian filters used. The half-peak radius of these filters are $10\cdot\sqrt{2}$, 20, $20\cdot\sqrt{2}$, and 40 pixels respectively.	65
3.8	This figure shows a subset of the elongated and oriented DoG filters used. The output of these filters was used to suppress the FPs due to linear structures such as blood vessels and ducts. .	67
3.9	This figure shows examples of images in which the linear structures were enhanced using the oriented DoG filters. Figures 3.9(a), 3.9(b) and 3.9(c) show the original images and figures 3.9(d), 3.9(e) and 3.9(f) show the respective enhanced images.	68
3.10	The set of 25 SFBs used for the detection of SM. To select the parameters of these SFBs an estimate of the joint pdf of the central mass region radius, length of spicules and number of spicules was obtained. This joint estimate of the pdf was then randomly sampled to select the parameters of the filter-banks.	71
3.11	FROC Curves: The performance of the detection algorithm is reported using FROC curves. The y-axis of the FROC curve corresponds to the sensitivity of the algorithm and the x-axis corresponds to the number of FPI per image. Figures 3.11(a) and 3.11(b) shows the FROC curve for the set of 50 spiculated masses and 46 architectural distortions from the DDSM database, respectively.	75

3.12	FROC Curves: Figure 3.12(a) shows the results on the set of images used by Karssemeijer <i>et al.</i> . If normal images are added to this set of spiculated masses, the number of false-positives per image will increase. The blue curve in figure 3.12(b) shows the FROC curve when the data-set contained images of both spiculated masses and 31 normal images.	76
3.13	FROC Curves: Figure 3.13(a) shows the results on the set of images used by Liu <i>et al.</i> . If normal images are added to this set of spiculated masses, the number of false-positives per image will increase. The blue curve in figure 3.13(b) shows the FROC curve when the data-set contained images of both spiculated masses and 19 normal images.	77
4.1	This figure shows a screenshot of the observer experiments carried out. The image in the center is the original image and the images on the left and right are the enhanced images obtained with the two enhancement methods. The observer is asked to choose which enhanced image they found most visually appealing. For the 2-AFC observer studies experiments the observers were not told from which method a particular enhanced image was generated and the images were shown in random order. . .	84
4.2	The only two images where most of the readers liked Method 1 better than Method 2. These were also the only two images where the radiologist found the enhanced image created with Method 1 more visually appealing.	85
4.3	Three images where most of the readers liked the images enhanced with Method 2 better than the images enhanced with Method 1.	86
5.1	This figure shows examples of the two extremes of segmentations. In Fig 5.1(b) the central mass region (dense object) of a spiculated mass is segmented. In Fig. 5.1(c) the spicules (linear structures) are outlined. The segmentations are represented by setting the pixels on or inside the boundary to one and zero otherwise.	115
5.2	A schematic explanation of the PCM algorithm. Let f and g represent the two segmentation images that are to be compared. Let $f(i, j)$ represent a particular pixel in f and $g(k, l)$ represent the corresponding pixel in g . For every pixel $f(i, j)$, the PCM algorithm tries to find its match within a neighborhood of radius r of the corresponding pixel $g(k, l)$	115

- 5.3 This figure shows the effect of rotation on similarity metrics. Note that the edges widths are fixed and no translation is applied. In Figs. 5.3(a), 5.3(b), and 5.3(c) the original image is displayed with solid lines. The locations of the lines of the image that was obtained after applying a rotation is shown in these images with dotted lines. In Figs. 5.3(a), 5.3(b), and 5.3(c) rotations of 1, 2, and 4 degrees were applied respectively. Figure 5.3(d) shows the effect of rotation on the intensity-based metrics whereas Fig. 5.3(e) compares the performance of the spatial-based metrics for different rotations. The performance of the Dice (intensity-based), PCM (spatial-based), and CW-SSIM are compared in Fig. 5.3(f). 116
- 5.4 This figure shows the effect of translation on similarity metrics. Note that the edges widths are fixed and no rotation is applied. In Figs. 5.4(a), 5.4(b), and 5.4(c) the original image is displayed with solid lines. The locations of the lines of the image that was obtained after applying a translation is shown in these images with dotted lines. In Figs. 5.4(a), 5.4(b), and 5.4(c) translations of 3.00, 5.10, and 7.28 units (Euclidean distance) were applied respectively. Figure 5.4(d) shows the effect of translation on the intensity-based metrics whereas Fig. 5.4(e) compares the performance of the spatial-based metrics for different translations. The performance of Dice (intensity-based), PCM (spatial-based), and CW-SSIM are compared in Fig. 5.4(f). 117
- 5.5 This figure shows the effect of edge width on similarity metrics. Note that a fixed rotation of 4 degrees was applied to the original image and no translation was applied in this simulation. In Fig. 5.5(a) the edge width was one pixel, whereas the simulated image was dilated by factors of 1 and 2 in Figs. 5.5(b) and 5.5(c) respectively. Figure 5.5(d) shows the effect of edge width on the intensity-based metrics whereas Fig. 5.5(e) compares the performance of the spatial-based metrics for different edge widths. The performance of Dice (intensity-based), PCM (spatial-based), and CW-SSIM are compared in Fig. 5.5(f). . . 118

5.6	This figure shows the effect of edge width on the similarity metrics as a function of rotation distortion. The original image was rotated from 0 to 4 degrees in increments of 0.4 degrees. The locations of the lines of the image that was obtained after applying a rotation are shown with dotted lines (for example, in Fig. 5.6(a), the original image is rotated by 4 degrees). The original image was compared to each of the rotated images and for each pair of images the similarity metrics were computed. To study the effect of the edge width, the image was dilated and the metrics were recalculated. Figures 5.6(c) and 5.6(e) show the simulated images after dilation by factors of 1 and 2 respectively and Figs. 5.6(d) and 5.6(f) show the similarity metrics as a function of rotation.	119
5.7	This figure shows the effect of edge width on the similarity metrics as a function of translation distortion. The original image was translated in the x and y directions by different amounts. The locations of the lines of the image that was obtained after applying spatial translations are shown with dotted lines (for example, in Fig. 5.7(a) a translation of 7 pixels to the right and 2 pixels to the bottom is applied). The original image was compared to each of the translated images and for each pair of images the similarity metrics were computed. To study the effect of the edge width on the metrics, the image was dilated and the metrics recalculated. Figures 5.7(c) and 5.7(e) show the simulated images after dilation by factors of 1 and 2 respectively and Figs. 5.7(d) and 5.7(f) show the similarity metrics as a function of translation.	120
5.8	Intra-observer agreement: This figure shows two sets of outlines of spicules created by a radiologist (R1) for four spiculated masses. The first column shows the original images and the first and second sets of outlines made by R1 are displayed in columns two and three. Column four shows the Dice, PCM, and CW-SSIM metrics for each pair of segmentations.	121
5.9	Inter-observer agreement: This figure shows the outlines of spicules created by two radiologists (R1 and R2) for four spiculated masses. The first column shows the original images and the second and third columns show the outlines made by R1 and R2 respectively. Column four shows the Dice, PCM, and CW-SSIM metrics for each pair of segmentations.	122

Chapter 1

Introduction

1.1 Motivation

The American Cancer Society estimates that 212,920 women will be diagnosed with breast cancer in the United States in 2006 [1]. Another 40,970 women will die of the disease. In the United States, breast cancer is the most common form of cancer among women and is the second leading cause of cancer deaths after lung cancer [1]. Women in the United States have about a 1 in 8 lifetime risk of developing invasive breast cancer [2, 3]. Early detection of breast cancer increases the survival rate and increases the treatment options. Screening mammography, x-ray imaging of the breast, is currently the most effective tool for early detection of breast cancer. Screening mammographic examinations are performed on asymptomatic woman to detect early, clinically unsuspected breast cancer. Two views of each breast are recorded; the craniocaudal (CC) view, which is a top-to-bottom view, and a mediolateral oblique (MLO) view, which is a side view taken at an angle. Examples of the MLO and CC views are shown in Fig. 1.1.

Radiologists visually search mammograms for specific abnormalities. Some of the important signs of breast cancer that radiologists look for are

clusters of microcalcifications, masses, and architectural distortions. A mass is defined as a space-occupying lesion seen in at least two different projections [4]. Masses are described by their shape and margin characteristics. Calcifications are tiny deposits of calcium, which appear as small bright spots on the mammogram. They are characterized by their type and distribution properties. An architectural distortion is defined as follows: “The normal architecture is distorted with no definite mass visible. This includes spiculations radiating from a point, and focal retraction or distortion of the edge of the parenchyma” [4]. A typical example of each of these abnormalities is shown in Fig. 1.2. Breast lesions are described and reported according to the Breast Imaging Reporting and Data System (BI-RADSTM) [4]. BI-RADSTM is a lexicon developed by the American College of Radiology (ACR). The BI-RADSTM lexicon for mammography includes descriptors such as the margin of a mass and the distribution of calcifications and it defines final assessment categories to describe the radiologists level of suspicion about the mammographic abnormality. It has been demonstrated that the BI-RADSTM final assessment rating is an indicator of the likelihood of malignancy [5]. If a suspicious abnormality is detected, a diagnostic mammographic examination is carried out to decide the future course of action required. Based on the level of suspicion of the abnormality following the diagnostic examination, a recommendation is made for routine follow-up, short-term follow-up, or biopsy.

Early detection via mammography increases breast cancer treatment options and the survival rate [6]. However, mammography is not perfect.

Detection of suspicious abnormalities is a repetitive and fatiguing task. For every thousand cases analyzed by a radiologist, only three to four are cancerous and thus an abnormality may be overlooked. As a result, radiologists fail to detect 10% to 30% of cancers [7–9]. Approximately two thirds of these false-negative results are due to missed lesions that are evident retrospectively [10]. Due to the considerable amount of overlap in the appearance of malignant and benign abnormalities, mammography has a positive predictive value (PPV) of less than 35% [11], where the PPV is defined as the percentage of lesions subjected to biopsy that were found to be cancer. Thus, a high proportion of biopsies are performed on benign lesions. Avoiding benign biopsies would spare women anxiety, discomfort, and expense.

Computer-aided detection (CAdE) systems have been developed to aid radiologists in detecting mammographic lesions that may indicate the presence of breast cancer. These systems act as a second reader and the final decision is made by the radiologist. Recent studies have also shown that CAdE detection systems, when used as an aid, have improved radiologists accuracy of detection of breast cancer [12–16]. However, recently there has been some controversy in their use and a recent study by Gur *et al.* found no increase in the number of cancers detected with current commercial CAdE systems [17].

Computer-aided diagnosis (CADx) systems for aiding in the decision between follow-up and biopsy are still in development. It is important to realize that mammographic image analysis is an extremely challenging task for a number of reasons. First, since the efficacy of CAdE/CADx systems can

have very serious implications, there is a need for near perfection. Second, the large variability in the appearance of abnormalities makes this a very difficult image analysis task. Finally, abnormalities are often occluded or hidden in dense breast tissue, which makes detection difficult.

1.2 Computer-Aided Detection of Mammographic Abnormalities

The goal of CAdE is to assist radiologists in locating abnormalities on the mammogram. A flowchart showing the different steps involved in detection algorithms is shown in Fig. 1.3.

The metrics used to report the performance of detection algorithms are sensitivity (Equation 1.1) and the number of false positives per image (FPI; Equation 1.1). A true positive mark is a mark made by the CAD system that corresponds to the location of a lesion. A false-positive mark is a mark made by the CAD system that does not correspond to the location of a lesion. A plot of sensitivity versus FPI is called a free-response receiver operating characteristic (FROC) plot and this is generally used to report the performance of the detection algorithm. An example of an FROC plot is shown in Fig. 1.4.

There is some disagreement regarding the manner in which detection results should be reported. While most authors report the performance in terms of the detection of any “actionable” objects, some report it terms of how many malignant masses were detected, since they believe that detection of malignant masses is most important. Whatever the methodology used, it is

necessary for researchers to clearly state the reporting method that has been adopted. Until recently, FROC analysis has been limited by the fact that the statistical analysis of FROC curves was less developed than that of traditional receiver operating characteristic (ROC). Major advances have recently been made in FROC analysis, particularly by Chakraborty and Berbaum [18]. However, despite the consistent use of evaluation methods in the literature, direct comparison of systems for detecting mammographic abnormalities is difficult because few studies have been reported on a common database.

$$Sensitivity = \frac{\text{Number of True-Positive Marks}}{\text{Number of Lesions}} \quad (1.1)$$

$$FPI = \frac{\text{Number of False-Positive Marks}}{\text{Number of Images}} \quad (1.2)$$

1.2.1 Detection of Masses

A mass is defined as a space-occupying lesion seen in at least two different projections [4]. Radiologists characterize masses by their shape and margin properties. A number of researchers have worked on methods for detecting masses in mammograms. Masses with spiculated margins have a very high likelihood of malignancy and thus some methods have been developed specifically for the detection of spiculated masses. A spiculated mass is characterized by lines radiating from the margins of a mass [4]. However, since not all malignant masses are spiculated, the detection of non-spiculated masses is also important. Most mass detection algorithms consist of two stages: (a) detection of suspicious regions on the mammogram and (b) classification of suspicious regions as mass or normal tissue. These are described in Sections

1.2.1.1 and 1.2.1.2, respectively.

1.2.1.1 Stage 1: Detection of Suspicious Regions

The first stage is designed to have a very high sensitivity and a large number of false positives are acceptable since they are expected to be removed in stage 2. Algorithms for stage 1 detection can generally be considered to be of two types, pixel based or region based [19].

Pixel-based Detection Methods: In pixel-based methods, features are extracted for each pixel and they are then classified as suspicious or normal. The terminology “pixel-based” is misleading since for every pixel, features are extracted from the local neighborhood of the pixel. This is followed by a classification step in which pixels are classified as suspicious or not. This may be done by simply applying a threshold to the feature image or by using sophisticated classification techniques. Finally, suspicious pixels are grouped together into regions, generally by collecting connected pixels.

It is important to emphasize that regions labeled as suspicious by the detection algorithms are not necessarily malignant. The classification of detected regions into malignant or benign categories is a different problem. A brief summary of pixel based mass detection methods follows. A number of detection methods have targeted particular subsets of masses. For example, some researchers have focused on the detection of spiculated masses because of their high likelihood of malignancy. The main idea behind the detection of spiculated masses is as that since spiculated masses are characterized by

spicules radiating in all directions, one should compute the edge orientations at each pixel. Thus, each pixel is represented by a feature vector that represents the strongest edge orientation at the pixel. The edge orientation can be computed in a variety of different ways.

Kegelmeyer *et al.* [20] developed a method to detect spiculated masses using a set of 5 features for each pixel. They used the standard deviation of a local edge orientation histogram (ALOE) and the output of four spatial filters which are a subset of Laws texture features. The idea of using the ALOE feature is that a normal mammogram exhibits a tissue structure that radiates in a particular orientation (from the nipple to the chest). A spiculated mass would change this trend and thus while normal tissue would have edge orientations in a particular direction whereas, in suspicious regions containing spiculated lesions, edges would exist in many different orientations. To detect this difference Kegelmeyer *et al.* [20] computed edge orientations in a window around each pixel and then generated a histogram of edge orientations. This idea is depicted in Fig. 1.5. The ALOE feature was then defined as the standard deviation of the bin heights of the histogram and is described by equation 1.3.

$$ALOE(\sigma_{ij}) = \frac{\sum_{n=0}^{255} (hist_{ij}(n) - \overline{hist}(i, j))^2}{255} \quad (1.3)$$

Where $hist_{ij}$ is the histogram of edge orientations in a window around the pixel located at (i, j) , and $\overline{hist}(i, j)$ is the average bin height of the histogram $hist_{ij}$. A binary decision tree was used to classify each pixel. The neighborhood size

for computing the ALOE was chosen to be 4 cm so that it would encompass all of the spiculated masses in the dataset.

Karssemeijer and te Brake [21] detected stellate distortions by a statistical analysis of a map of pixel orientations. The orientation at each pixel was computed from the response of three filter kernels, which are second-order, directional derivatives of a Gaussian kernel in the directions $(0, \pi/3, 2\pi/3)$. These filters form a non-orthogonal basis and are shown in Fig. 1.6. They used the relation that at a particular scale (σ), the output at any orientation $W_\sigma(\theta)$ can be expressed as a weighted sum of the responses of the filters. This is described in Equation 1.4, where $W_\sigma(0)$, $W_\sigma(\pi/3)$ and $W_\sigma(2 * \pi/3)$ are the responses of the three filters. This relation was used to determine the orientation at each pixel and two features for each pixel were derived by a statistical analysis of these pixel orientation maps. The pixels were then classified as suspicious or normal. To account for the range of sizes of spiculations in their dataset, edge orientations were computed at three spatial scales ($\sigma = 1, 2, 3$) and the one with the maximum magnitude was used. We note that this is equivalent to choosing local neighborhoods of varying sizes.

$$\begin{aligned} W_\sigma(\theta) = & \frac{1}{3}[1 + 2 \cdot \cos(2\theta)](W_\sigma(0)) \\ & + \frac{1}{3}[1 - \cos(2\theta) + \sqrt{3}\sin(2\theta)](W_\sigma(\pi/3)) \\ & + \frac{1}{3}[1 - \cos(2\theta) - \sqrt{3}\sin(2\theta)](W_\sigma(2\pi/3)) \end{aligned} \quad (1.4)$$

Liu *et al.* [22] point out that in general, it is difficult to estimate the size of the neighborhood that should be used to compute the local features

of spiculated masses. Small masses may be missed if the neighborhood is too large and parts of large masses may be missed if the neighborhood is too small. To address this problem Liu *et al.* [22] developed a multiresolution algorithm for the detection of spiculated masses. They generated a multiresolution representation of a mammogram using the discrete wavelet transform. They extracted four features at each resolution for each pixel. One of the features they used was the ALOE feature described in Equation 1.3. Pixels were then classified using a binary classification tree. The detection was carried out in a top-down manner from the coarsest resolution to the finer resolutions. If a positive detection was made and a pixel was classified as abnormal, no feature extraction and detection were needed at the corresponding pixels at all finer resolutions. This approach reduced the number of pixels to be classified.

In all of the three methods described above, the focus was on developing sophisticated stage 1 detection techniques. These methods used very simple techniques for the stage 2 task. For example, Karssemeijer and te Brake [21] grouped suspicious regions and discarded regions that were smaller than 500 pixels. Other researchers have not restricted their efforts to the detection of spiculated masses since many malignant masses are not spiculated. Li *et al.* [19] developed a two-step process for detection of masses. In the first step, adaptive gray-level thresholding was used to obtain an initial segmentation of suspicious regions. The segmentation was iteratively improved using a multiresolution Markov random field (MRF) based segmentation method. The algorithm was first applied at the coarsest resolution and the output was refined

at the next finer resolution. This strategy helps to reduce the computational complexity as mentioned above.

In the second stage, a fuzzy binary decision tree was used to classify the segmented regions as masses or normal tissue using features based on shape, region size, and contrast. Matsubara *et al.* [23] developed an adaptive thresholding technique for the detection of masses. They used histogram analysis techniques to divide mammograms into three categories ranging from fatty to dense tissue. Potential masses were detected using multiple threshold values based on the category of the mammogram. A number of features such as circularity, area, and standard deviation were used to reduce the number of false positives. Li *et al.* [24] developed a method for lesion site selection using morphologic enhancement and stochastic modelbased segmentation technique. A finite generalized Gaussian mixture distribution was used to model histograms of mammograms. The expectation maximization algorithm [25] was used to determine the parameters of the model. The segmentation was achieved by classifying pixels using a new Bayesian relaxation labeling technique. An underlying motivation for this technique was that it could incorporate neighborhood information into the classification process and that this would help improve the process. They argued that for the purpose of lesion site selection, sensitivity should be the sole criterion for evaluation and thus did not incorporate a false-positive detection step. The primary advantage of using pixel-based methods is that one has a large number of samples to train a classifier. However, this class of methods also has inherent disad-

vantages. It does not take into account the spatial arrangement of the pixels, which is a very important factor to discriminate masses from normal tissue. A different set of features would be required to describe different mass types. It is computationally intensive and hence, most pixel-based methods must sub-sample images before detection. The advantage of having many pixels per image available for use in training supervised learning methods should not be overstated. There are two problems regarding the use of multiple pixels. First, pixels at the periphery of a mass and at the center of the mass belong to the same class, but are not always homogeneous and maybe represented by different feature values. This is a major limitation as, ideally, one would want samples of a particular class to possess similar feature values. Second, multiple pixels from a single mass represent only one particular lesion example. This does not eliminate the need for a comprehensive database containing masses that encompasses the range of natural variability of masses.

Region-based Detection Methods: In region based detection methods, regions of interest are first extracted by a segmentation or filtering technique. Features are then extracted for each region and the region is classified as suspicious or otherwise. These features are designed to describe important diagnostic information like shape and texture of the extracted regions. A number of these methods are based on the idea of matched filtering. In these approaches, the image is filtered with a filter that is used as a model for a mass. The idea is that the output of the filtered image will be high near the center of the tumor masses. Often the N largest outputs are selected as possible suspicious regions.

This is followed by the extraction of ROIs around the N largest peaks. Features are extracted from the ROI, and the ROIs are classified as containing a mass or normal tissue. Here lies the main difference between pixel and region-based detection methods. In the pixel-based methods, features were extracted for each pixel, whereas in the region-based methods, features are extracted for each region. A brief description of the region-based methods that used a matched filtering approach is given below.

Kobatake *et al.* [26] modeled masses as rounded convex regions and based on this idea, developed an “iris filter” to enhance and detect masses. The iris filter was applied to a gradient image that was generated by Prewitt-type operators. The output of the filter was computed by measuring the average convergence of the gradient over the region of support of the filter. The peaks of the output of the filter were selected as centers of tumor candidates. The filter was then reapplied locally to detect the boundaries of candidate masses. Finally, texture features were computed from the candidates and were used to reduce false-positives. The authors showed that one of the advantages of using this filter was that the output of the filter would be constant regardless of the contrast between a rounded convex region and the background.

Petrick *et al.* [27] developed a two-stage algorithm for the enhancement of suspicious objects. In the first stage, they proposed an adaptive density-weighted contrast-enhancement (DWCE) filter to enhance objects and suppress background structures. The central idea of this filtering technique was that it used the density value of each pixel to weight its local contrast. In the

first stage, the DWCE filter and a simple edge detector (Laplacian of Gaussian) were used to extract ROIs containing potential masses. In the second stage, the DWCE was reapplied to the ROI. Finally, to reduce the number of false positives, they used a set of texture features for classifying detected objects as masses or normal. They further improved the detection algorithm by adding an object-based region growing algorithm [28].

Polakowski *et al.* [29] used a single difference of Gaussian (DoG) filter to detect masses. The DoG filter was designed to match masses that were approximately 1 cm in diameter. ROIs were selected from the filtered image. They used nine features based on size, contrast, circularity and Laws texture features to reduce the number of false positives and to then classify ROIs as malignant or normal. The DoG filter, which is a band-pass filter, has been used by several researchers for the preliminary task of detection of potential masses in an image. The DoG filter must be matched to the size of the mass. Since the size of masses varies from a few millimeters to several centimeters [30], a number of DoG filters would be required, which would increase the computational complexity. Since the size of a potential mass is not known *a priori*, several researchers have used multiscale region-based methods for the detection of masses.

Brzakovic *et al.* [31] use a two-stage multiresolution approach for detection of masses. First, they identified suspicious ROIs using Gaussian pyramids and a pyramid linking technique based on the intensity of edge links. Edges were linked across various levels of resolution. This was followed by a classifi-

cation stage, where the ROIs were classified as malignant, benign, or normal on the basis of features like shape descriptors, edge descriptors, and area.

Qian *et al.* [32] developed a multi-resolution and multi-orientation wavelet transform for the detection of masses and spiculation analysis. They observed that traditional wavelet transforms cannot extract directional information, which is crucial for a spiculation detection task and thus, they introduced a directional wavelet transform. Figure 1.7 shows the partitioning of the frequency domain with the directional wavelet transform. We note that in comparison, a conventional wavelet transform would produce a rectangular partitioning of the frequency domain. An input image was decomposed into two output images using the directional wavelet transform. One was a smoothed version of the original image and was used to segment the boundary of the mass. The second contained the high-frequency information and was used for directional feature extraction. The key ideas of the method were that at coarser resolutions, features such as the central mass region can be easily detected, whereas at finer resolutions, detailed directional features such as spicules can be localized.

As was the case for pixel-based methods, some researchers have developed region-based methods that are focused on the detection of masses with particular margin characteristics, such as circumscribed or spiculated masses. Lai *et al.* [33] developed a simple template matching algorithm to detect circumscribed masses only. They enhanced images using a modified median filtering technique to remove background noise. To cope with variations in the

size of masses various templates with radii ranging from three to 14 pixels were used. To measure the similarity between a potential mass and the template, the authors chose the normalized cross-correlation as a similarity metric. This particular metric was chosen since it is invariant to the size of the template and the average brightness of the image. They also developed two features to reduce the number of false positives detected.

Groshong and Kegelmeyer [34] used the circular Hough transform for the detection of circumscribed lesions. The Hough domain, for circular objects consists of three parameters (x , y , and r) corresponding to the x and y centers and radius (r) of the object. Thus, a point in the three dimensional Hough domain maps to a circle in the image domain. They computed an edge image using a Canny operator and selected a subset of the edges based on length and intensity. This subset of edges was the input to a circular Hough transform. The radius parameter search space ranged from 3 to 30 mm to account for masses of different sizes. Two features were extracted from the Hough domain for each pixel and ultimately these were classified as either belonging to a mass or normal tissue.

Zhang *et al.* [35] noted that the presence of spiculated lesions led to changes in the local mammographic texture. They proposed that such a change could be detected in the Hough domain, which is computed using the Hough transform. They partitioned an image into overlapping ROIs and computed the Hough transform for each ROI. The Hough domain of each ROI was thresholded to detect local changes in mammographic texture and to

determine the presence or absence of a spiculated mass. Region-based methods have a number of advantages. In contrast to pixel-based methods, region-based detection takes into account the spatial information. Also, the features are directly correlated to important diagnostic information like the shape and margin of extracted regions. They are computationally less intensive than pixel-based methods. The main disadvantage is that if a classifier is used, there are fewer samples for training the classifier as compared to the pixel based methods.

To conclude this section, we note that masses can have a range of sizes. Thus, a major limitation of both pixel-based and region-based methods is that the analysis is not done over a continuous range of scales. Cancerous lesions are stochastic biologic phenomena that manifest in images as having various structures occurring at different sizes and over ranges of spatial scales. For example, masses occupy definite regions; this region occupancy can be approached at a coarse scale of description or processing. However, the boundaries of masses require a more localized approach, although the sharpness, and hence the scales of interpretation of the lesion boundaries, can vary considerably. Moreover, the spiculations that are associated with many cancerous lesions occur with different widths, lengths, and densities, which suggests that their characterization will require analysis over scales.

1.2.1.2 Stage 2: Classification of Suspicious Regions as Mass or Normal Tissue

A number of researchers have focused solely on the second stage of detection in which suspicious regions are classified as mass or normal tissue. The purpose of the second stage is to reduce the number of false positives that were produced at the end of the first stage. A brief summary of stage two methods follows.

Researchers have used texture features to discriminate between mass and normal tissue. Sahiner *et al.* [36] proposed a convolution neural network for this task. They extracted texture features from the ROIs. Wei *et al.* [37] developed a classifier using texture features and linear discriminant analysis for this task. They computed multiresolution texture features from spatial gray-level dependence matrices. Wei *et al.* [38] also investigated the use of global and local multiresolution texture features for this task and for reducing the number of false-positive detections on a set of manually extracted ROI.

Radiologists use a number of image characteristics to discriminate between masses and normal tissue and researchers have attempted to emulate that process. te Brake *et al.* [39] defined a number of features to discriminate between lesions and normal tissue that were designed to capture image characteristics like intensity, iso-density, location, and contrast. Kupinski and Giger [40] studied a regularized neural network for this task. Masses were detected using the bilateral subtraction scheme. Features based on geometry intensity and the gradients of potential lesions were extracted. They also evaluated

the effectiveness to minimize over training. Mutual information and a subregion hotelling observer have also been tested for this classification problem. Tourassi *et al.* [41] developed a template-matching technique for this problem. Each ROI in the database served as a template and mutual information was used a similarity metric to decide if a query ROI contained a mass. Baydush *et al.* [42] proposed a subregion hotelling observer for detecting whether a given ROI contained a mass or not.

1.2.1.3 Methods for the detection of architectural distortions

In comparison to the detection of masses there have been few studies which have focused on the detection of architectural distortions (AD). This section briefly reviews recent methods developed specifically for the detection of AD. Hara *et al.* [43] developed a method to detect AD based on the concentration of the mammary gland. They used dynamic range compression as a preprocessing step. The mean curvature, shape index and curvedness were used to extract the mammary gland.

Rangayyan and Ayres [44] developed a technique for the detection of AD which is based on the analysis of oriented texture through a linear phase portrait model. To reduce the number of false-positives they constrained the shape of the general phase portrait model.

Elmaghraby *et al.* [45] developed a scheme for the detection of AD that focuses on the morphological effects of breast parenchyma due to cancer infiltration. Their hypothesis was that ADs evolve in concentric layers around a focal region. In their approach they first select potential seeds by identifying

areas of increased concentric activity. Finally, only those seeds with a high probability of malignancy are marked.

1.2.2 Commercial Computer-Aided Detection Systems

Three FDA-approved commercially available CADe systems have been developed to aid radiologists in detecting mammographic abnormalities. Currently, there are no FDA-approved systems for CADx. A brief description of each of these systems is given below.

1.2.2.1 R2 Technology, Inc.

R2 Technology's Image CheckerTM was the first commercial mammographic CADe system approved by the FDA [46]. This device is designed to search for signs that may be associated with breast cancer. Masses are marked with an asterisk, while microcalcification clusters are marked with a triangle. The detection accuracy of calcifications is reported as 98.5% sensitivity at 0.74 false positives per case (set of four images). The detection accuracy of masses is reported as 85.7% at 1.32 false-positive marks per case.

1.2.2.2 Intelligent Systems Software, Inc.

The FDA approved the Intelligent System Software Inc. (ISSI) CADe system MammoReaderTM in 2002. MammoReaderTM was designed to detect primary signs of breast cancer in mammogram images including microcalcification clusters, well- and ill-defined masses, spiculated lesions, architectural distortions, and asymmetric densities. Masses are marked with crosshairs and

microcalcification clusters with outlines [47]. The reported overall sensitivity is 89.3% (91.0% in cases where microcalcifications are the only sign of cancer and 87.4% in the remaining cases where malignant masses are present). The system made 1.53 true-positive marks and 2.32 false positive marks per case among cancer cases and 3.32 false positive marks among cases without cancer.

1.2.2.3 CADx Medical Systems

CADx Medical Systems was the third company to receive approval for a mammographic CAdE system called Second- LookTM [48]. SecondLookTM was designed to mark areas of a mammogram that are indicative of cancer. It marks masses with circles and microcalcification clusters with rectangles. The sensitivity of the system is reported to be 85% for screening-detected cancers (combination of masses and microcalcification clusters). Additionally, it identified cancer locations in 26.2% of mammograms acquired within 24 months before cancer diagnosis. CADx did not report SecondLook'sTM false-positive rate.

1.2.3 Independent Studies of Commercial Computer-Aided Detection Systems

Several large-scale independent trials of the R2 Image CheckerTM system have been conducted to test the performance of this system in a clinical setting. In a study conducted by Vyborny *et al.* [15], it was shown that the R2 Image CheckerTM system detected 86% of the spiculated masses at 0.24 FPI on a dataset of 375 images whereas it had a detection sensitivity of 53%

for nonspiculated masses. All of these masses were given a subtlety rating of subtle, medium, or obvious by three radiologists. It is important to note that of the 375 clearly spiculated masses, 271 were classified as “obvious”, 73 had a subtlety rating of “medium”, and only 31 had a subtlety rating of “subtle”. While the R2 Image CheckerTM system detected 94% of the obvious spiculated masses, it detected only 70% of the medium spiculated masses and 52% of the subtle spiculated masses. Though the overall results (86% at 0.24 FPI) for the detection of spiculated masses are impressive, it is important to note that the study used a large number of obvious masses and a much smaller number of subtle masses. Thus, there is still room for improvement, even in the detection of spiculated masses.

Freer and Ulissey [14] tested the performance of the R2 Image Checker system on more than 12,860 patients in a community breast center. For the first 20,624 radiographs, they observed that 14,214 computer cues or marks were made by the CAdE system. Of these, 13,846 marks (97.4%) were dismissed by the radiologist as false positives. This corresponds to a false-positive rate of 0.671 FPI. The CAdE system detected 67% (18 of 27) malignant masses and 100% of the clustered calcifications (22 of 22). The authors argue that dismissing the large number of false-positive marks was easy for a radiologist to do [14]. However, another study claims that dismissing false-positive cues can be difficult [49]. This study clearly showed that the R2 Image Checker systemTM is better at the detection of calcifications than at the detection of masses.

Baker *et al.* [50] studied the performance of two CAdE systems for the detection of architectural distortions on a set of 80 images. They observed that the R2 Image CheckerTM system had a sensitivity of 38% at 0.7 FPI while the CADx SecondLookTM system had a sensitivity of 21% at 1.27 FPI. They concluded that the sensitivity of current systems for the detection of architectural distortions is very low and that considerable improvements are needed for this detection task. On the basis of these studies, radiologists tend to trust the calcification cues more than the mass cues. This is also documented in the literature and prominent radiologists like C. J. D’Orsi published papers saying that “I would initially use only the calcification prompt and feel extremely comfortable that I have not missed any substantial calcifications when no cues for calcium are present” [51]. Thus, there is room for improvement in the detection accuracy of architectural distortions.

All of the commercially available CAdE systems perform much better at detecting calcifications than at detecting masses or architectural distortions. One also cannot make a direct comparison of these systems as there has been no clinical study that compares the performance of these systems on the same set of cases. Bowyer *et al.* [52] have argued that the fundamental advantage of the pixel level detection step is that a large amount of data is available for classifier training. They claim that extracting features from each pixel provides thousands of training samples to characterize normal/abnormal tissue. However, despite the large number of training samples and more than 25 years of research in mass and spiculated mass detection methods, there is significant

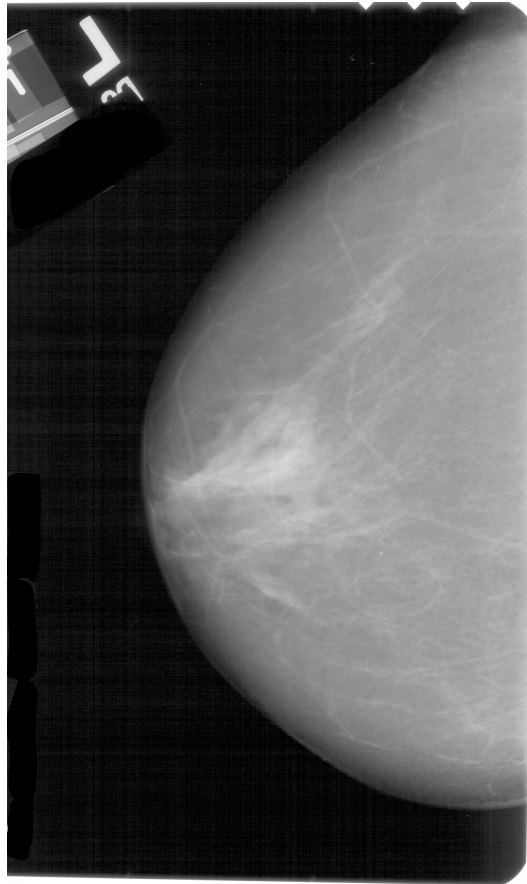
room for improvement in the detection accuracy of these algorithms. The detection sensitivity which ranges from 75% to 89% is much lower than the detection sensitivity of calcifications which is at about 98%. This shows that there is a pressing need for a new detection philosophy.

1.3 Overview of dissertation

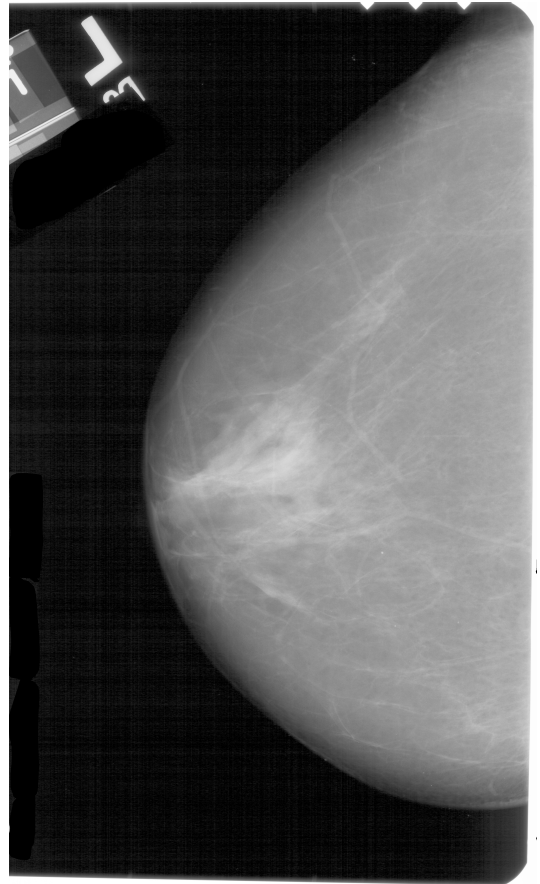
In this dissertation, a new *evidence-based* method for the detection of spiculated masses and architectural distortions was developed. By evidence based, we mean that we use the statistics of the physical characteristics of these abnormalities to determine the parameters of the detection algorithm. Towards this goal, it was shown that the properties of these lesions can be measured reliably and we have created the first database of the physical properties of these lesions.

The detection algorithm uses the basic definition of these lesions consists of two steps, an enhancement step followed by a filtering step. In the first step, a new technique for the enhancement of spiculations is developed, in which a linear filter is applied to the Radon transform of the image. The goal of the second step is to detect the spatial locations where these enhanced linear structures converge. This is achieved by filtering the enhanced image with a new class of linear image filters called Spiculation Filters. We have invented these filters specifically for detecting spiculated masses and architectural distortions that are marked by converging lines or spiculations. These filters are highly specific narrowband filters, which are designed to match the expected

structures of these abnormalities and form a new class of wavelet-type filter-banks derived from optimal theories of filtering. A key aspect of this work is that each parameter of the filter has been designed to capture the variation in physical characteristics of spiculated masses and architectural distortions and that all of the parameters of the detection algorithm are determined by the physical measurements. Techniques to reduce the number of false-positives were developed and the results of the detection algorithm are presented in the form of FROC curves.



(a) Craniocaudal (CC) view



(b) Mediolateral oblique (MLO) view

Figure 1.1: In screening mammography two views of each breast are recorded; the craniocaudal (CC) view (left), which is a top-to-bottom view, and a mediolateral oblique (MLO) view (right), which is a side view taken at an angle. The images were obtained from Digital Database for Screening Mammography (DDSM), [62]

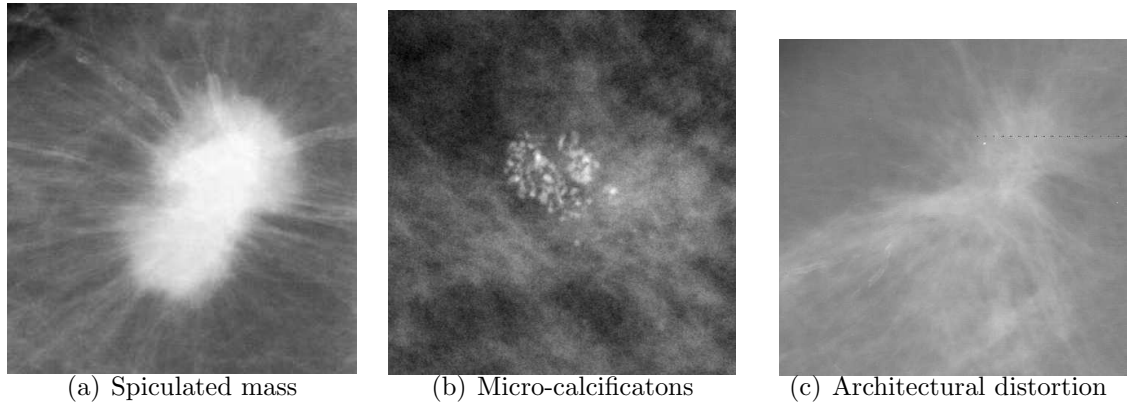


Figure 1.2: Examples of a spiculated mass (left), cluster of microcalcifications (center), and architectural distortion (right). The images were obtained from the DDSM, [62]

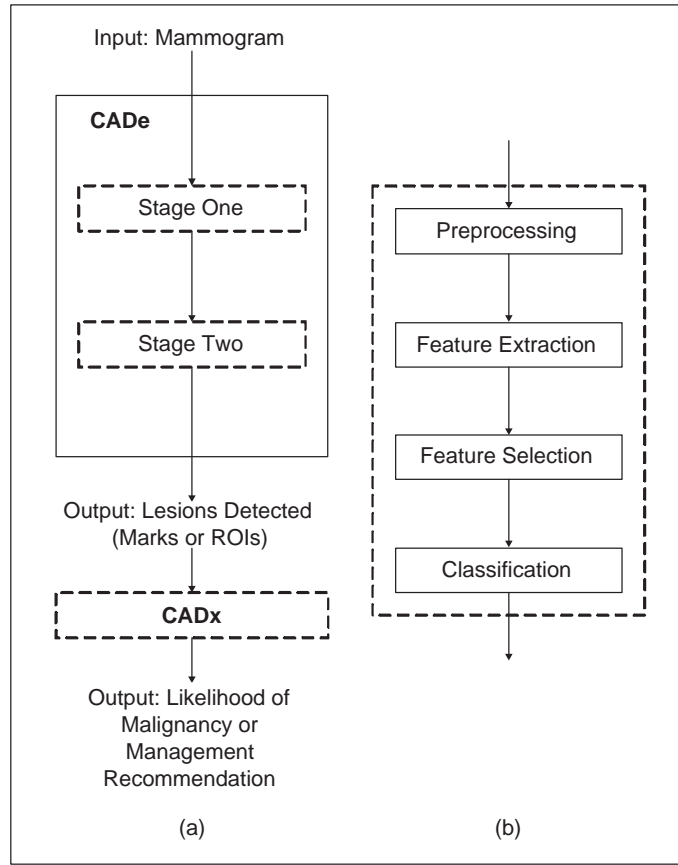


Figure 1.3: A flowchart showing the main steps involved in the computer-aided detection (CAdE) and computer-aided diagnosis (CAdx) of mammographic abnormalities. Most detection algorithms consist of two stages. In stage 1, the aim is to detect suspicious lesions at a high sensitivity. In stage 2, the aim is to reduce the number of false positives without decreasing the sensitivity drastically. The steps that are involved in designing algorithms for both stages are shown in (b). We note that in some approaches some of the steps may involve very simple methods or be skipped entirely. For example, in stage 1, the classification step often is a simple size criteria (i.e., if the size of potential lesion is suspicious only if its size is greater than N pixels). Most diagnosis algorithms (CAdx) begin with a region of interest (ROI) containing the abnormality. Again, the steps typically involved in design such a system are shown in (b). The output of a CAdx system may be the likelihood of malignancy or a management recommendation. Different research groups have worked on different components of the problem and human interaction may occur at various stages. For example, many CAdx algorithms start with manually segmented ROIs.

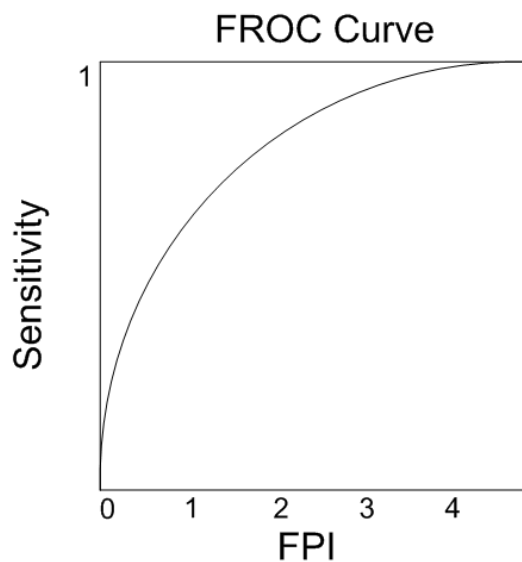


Figure 1.4: Free-response receiver operating characteristic (FROC) curve: In an FROC curve, sensitivity is plotted on the y-axis and the number of FPI is plotted along the x-axis. FROC curves are used to report the performance of detection studies.

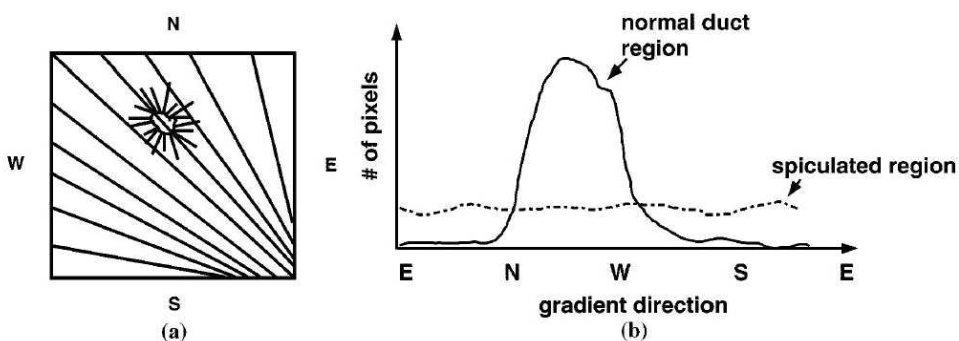


Figure 1.5: (a) Directions of spicules of a spiculated lesion differ from the directions of normal linear markings in a mammogram and the (b) standard deviation of the gradient orientation histogram differentiates the area near a spiculated lesion from normal. (©2006 IEEE).

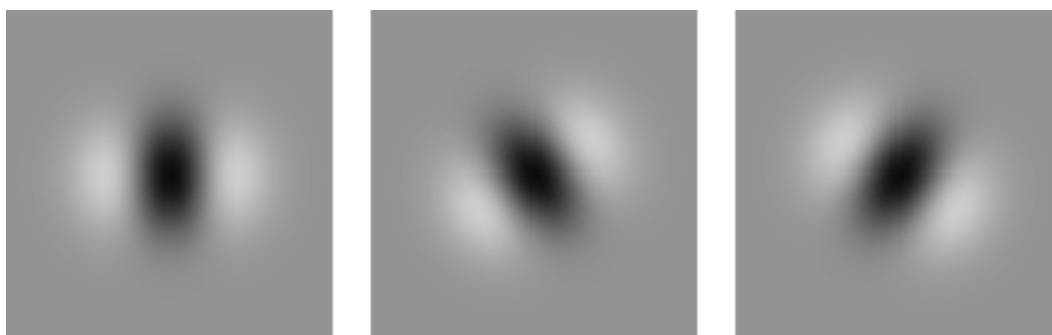


Figure 1.6: Three-directional second-order Gaussian derivatives used for estimation of line orientation. The figure was obtained from Prof. Nico Karssemeijer (©2006 IEEE).

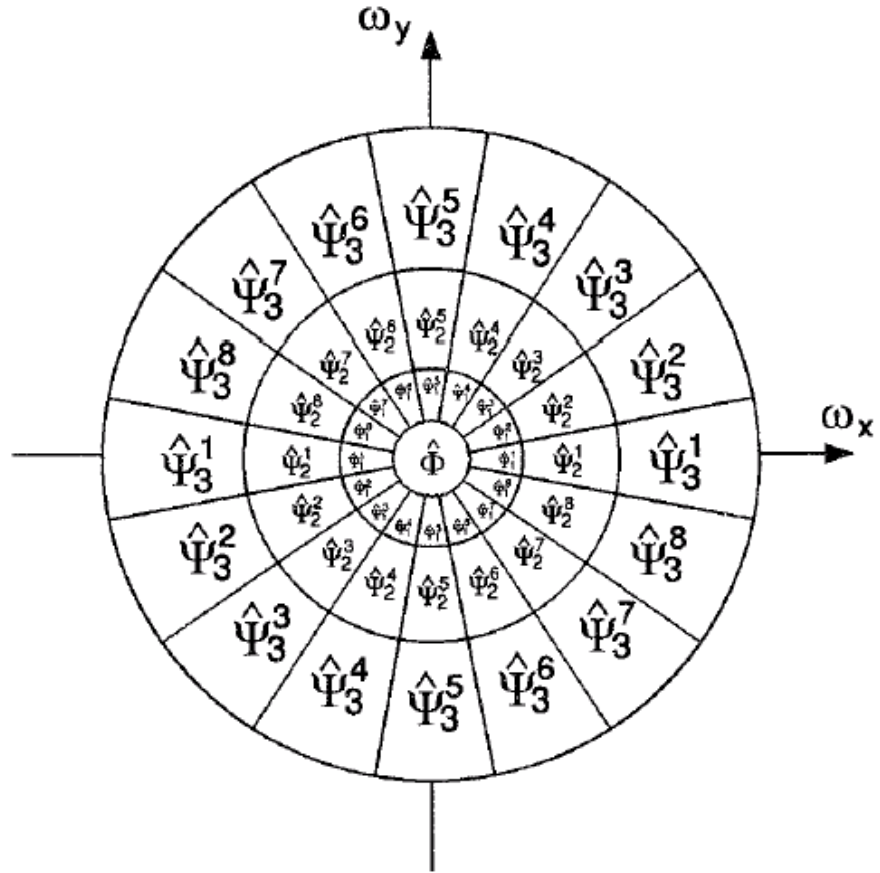


Figure 1.7: Partitioning of the frequency domain achieved with the directional wavelet transform (©2004 IEEE).

Chapter 2

Measurement Studies: The reliability of measuring physical characteristics of spiculated masses on mammography

2.1 Introduction

The central theme of this dissertation is to develop an *evidence-based* technique for the detection of spiculated lesions. To achieve this goal, measurements of the salient physical parameters of these lesions were collected. This chapter, describes the process of collecting measurements and the statistical methodology used to analyze the measurement data. The goal of this analysis was to assess the reliability of measurements of the physical characteristics of spiculated masses on mammography.

It is widely acknowledged that current CAdE systems detect microcalcifications more accurately than they detect masses, including spiculated masses. One reason for this is that calcifications are typically much denser than the surrounding tissue, whereas there is less contrast between masses and the parenchyma. Moreover, from an image processing perspective, calcifications are easier to detect because they can be simply modeled as impulse functions. In comparison, spiculated masses are difficult to model because of

the great variability in their physical characteristics. The lack of statistical information on the physical properties of spiculated masses makes it difficult for engineers to create mathematical models of these abnormalities. For instance, there is no quantitative record of the physical characteristics of spiculated masses, such as the typical length of spicules. This information would be beneficial for the design of CAdE algorithms [e.g., [53]], even though radiologists may not consciously use such information in detecting or characterizing lesions.

All radiological measurements are subject to inter- and intra-observer variability. A number of statistical methods are available to quantify inter- and intra-observer agreement. The Bland-Altman technique, intraclass correlation coefficient (*ICC*), Kappa statistic, and regression analysis are some of the most frequently used methods. While we are unaware of any studies that have focused on the reliability of measurements of mammographic lesions, several studies have assessed the observer variability of rating data, as opposed to measurement data, in mammographic interpretation. For example, considerable inter-observer variability has been reported in describing mammographic masses using the BI-RADS lexicon ([50, 54]). By comparison, many studies have evaluated the inter- and intra-observer variability of measurements in non-mammography medical imaging applications (e.g., [55–61]).

This chapter presents the results of a study in which two experienced radiologists measured the parameters of spiculated masses on mammography. It is demonstrated that the physical properties of spiculated masses can be

measured reliably on mammography and that the observer variability for this task is comparable to what has been reported in other medical imaging measurement studies.

2.2 Materials and Methods

2.2.1 Data Set

The images used in this measurement study were obtained from the Digital Database for Screening Mammography (DDSM), <http://marathon.csee.usf.edu/Mammography/Database.html> [62]. The DDSM is the largest publicly available data set of digitized mammograms. The entire database consists of 2620 cases and each case consists of four mammograms: a CC and MLO view of each breast. The mammograms were obtained from three institutions [62]. Along with the digitized mammograms, the DDSM contains “boundary” files of the abnormalities. The outlines of the abnormalities, as indicated by a radiologist, are stored in ‘chain code’ in these files. From this ‘chain code’ borders of the abnormalities can be reconstructed. In this study, observers primarily worked with a region of interest (ROI) from each image, though the full mammogram was always available to them. The ROI was defined such that the central mass and all spicules were visible. In particular, the ROI was taken as the smallest rectangle in which the boundary specified in the DDSM database could be inscribed, plus 500 pixels in each direction.

For this study, the MLO views of cases of spiculated masses were randomly selected from the DDSM. Cases were selected from a single scan-

	Total No. of cases	No. of malignant cases	Minimum Density	Average Density	Maximum Density
Training Set	12	10	1	2	3
Measurement Set	21	21	1	2	3

Table 2.1: Properties of the two sets of images used in this study. A set of 12 cases was used for observer training and measurements were collected using a second, distinct set of 21 cases.

ner, and we confirmed that a range of density ratings, subtlety ratings, and pathologies were represented by the sample. A set of 12 cases was used for observer training and measurements were collected using a second, distinct set of 21 cases. The characteristics of the measurement set are summarized in Table 2.1. A list of the DDSM cases numbers is given in Appendix A and the ROI images used in this study are available on our website www.bme.utexas.edu/research/informatics.

2.2.2 Observer Training and Measurement Protocol

The physical measurements were made by two experienced radiologists. For the rest of the paper, we will refer to these radiologists as R1 and R2. Radiologist R1 was the senior radiologist and had more experience in breast imaging. R1 was trained as a breast-imaging fellow for 1 year and has been reading mammograms since 1990. Currently, radiologist R1 reads 7,000 of mammograms per year. Radiologist R2 was trained as a breast-imaging fellow for 1 year and has been reading mammograms since 1994. Currently, radiol-

ogist R2 reads 3,000 of mammograms per year. We used the ROI Manager plugin of NIH ImageJ (<http://rsb.info.nih.gov/nih-image/>) to enable the radiologists to measure physical properties of spiculated lesions on mammograms (Figure 2.1). Using a straight line tool, the radiologists marked the principal axes of the central mass, the width of each spicule at its base where it meets the mass, and traced each spicule in order to measure its length. Since the resolution of the images was known, the pixel measurements were converted into physically meaningful quantities (e.g., mm). In addition, the radiologists counted the spicules associated with each lesion. The measurements were made on ROIs, but the radiologists were allowed to view the full mammogram at any time and could adjust the display as desired (e.g., zoom). The images were displayed on a standard laptop computer in a darkened room and the radiologists were allowed unlimited time for the measurement task. All of the images with the radiologists' markings overlaid are available on our website www.bme.utexas.edu/research/informatics.

Measuring spiculated lesions is not part of routine clinical practice. Thus, we conducted a training stage in which the radiologists discussed the results of measurements made independently on a training set of images. Each radiologist independently measured the properties of a training set of 12 spiculated masses. Their markings were overlaid on the original ROIs (Figure 2.2(a)) and they discussed areas of agreement and disagreement in their measurements. Following the training phase, the two radiologists independently measured the properties of 21 images of spiculated masses. There was no

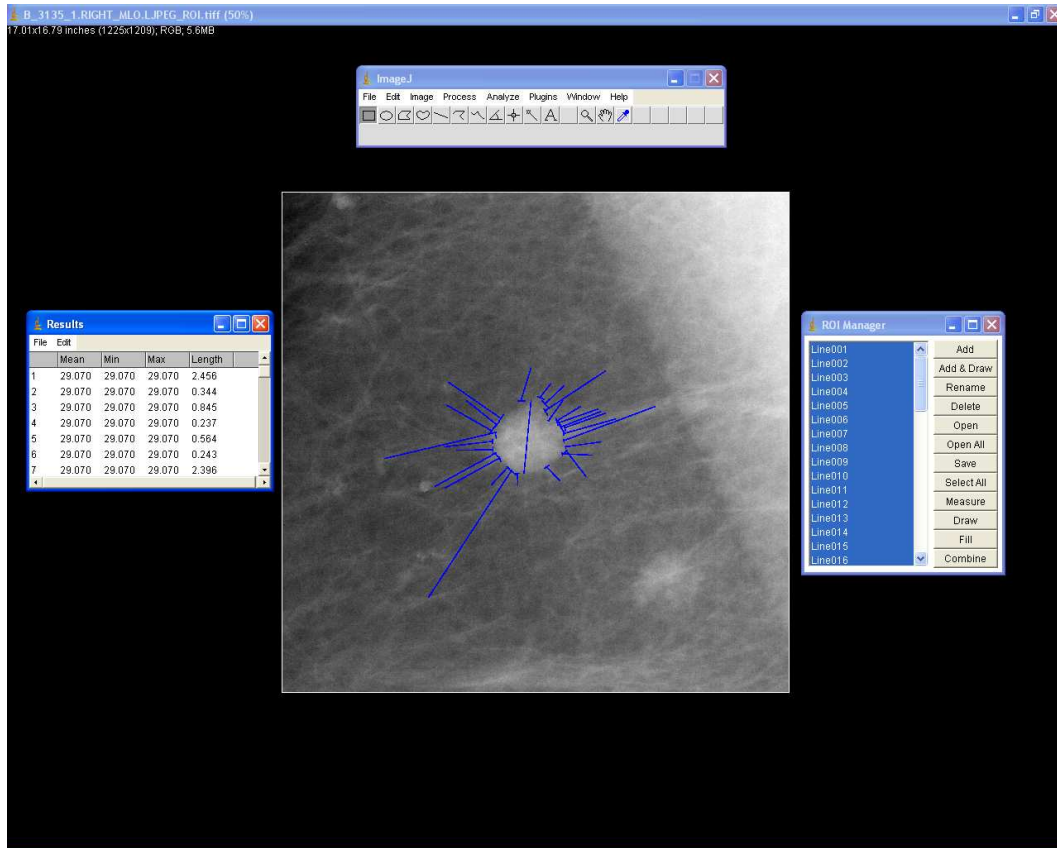
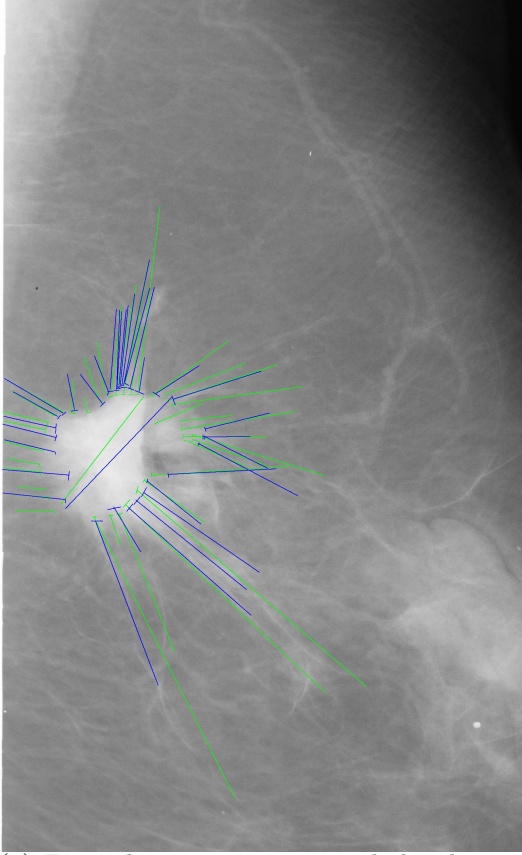


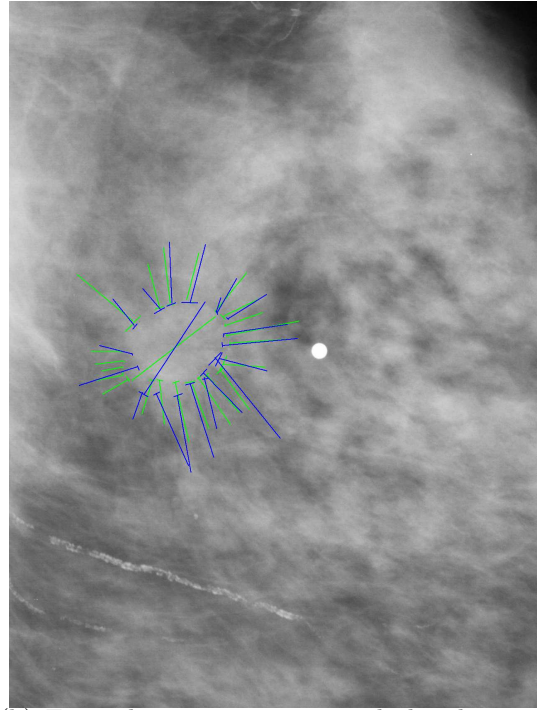
Figure 2.1: NIH ImageJ Interface for obtaining measurements of key characteristics of spiculated masses.

overlap between the training and the measurement sets. Because of time and scheduling constraints, these measurements were carried out in two sessions. In the first session (a few weeks after the training session), the properties of 12 images were measured and in the second session (a few months after the training session) the properties of the remaining 9 images were analyzed. To assess the intra-observer variability, one radiologist (R1) re-measured the first

set of 12 images after an interval of 5 months. Thus, a total of 21 images were used for the analysis of the inter-observer agreement and a set of 12 images was used to compute the intra-observer agreement.



(a) Example measurements made by the two radiologists during the training phase



(b) Example measurements made by the two radiologists during the measurement phase

Figure 2.2: Examples of the measurements made by radiologists R1 and R2 before and after the training stage. The measurements made by R1 are shown in blue and those made by R2 are shown in green.

2.2.3 Statistical Analysis

We believe that it is important to assess the degree of agreement using multiple statistical methods. This view is also shared by Luiz *et al.* [63] who noted that for the analysis of measurement studies it is desirable to report the degree of agreement using multiple statistical methods as no method is fool-proof and each has its own limitations. The degree of agreement between the measurements of radiologists R1 and R2 was evaluated using a hypothesis test for equivalence, the intraclass correlation (*ICC*) coefficient [64], and Bland-Altman statistics [65, 66]. In testing for equivalence, the null hypothesis is that the measurements of the two radiologists are not equivalent and the alternative hypothesis is that they are equivalent [67]. Note that the more familiar paired t-test for a null hypothesis of equal values vs. an alternative hypothesis of not equal values is not an appropriate test for establishing equivalence. Failing to reject a null hypothesis does not prove that the null hypothesis is correct; in particular, a failure to reject the null hypothesis can arise from a lack of power. Thus, a hypothesis test specifically intended for assessing equivalence was used.

The test statistic (t) for assessing equivalence is:

$$t = \frac{\bar{x} \pm \delta}{s} \quad (2.1)$$

Where \bar{x} and s are the mean and standard deviation respectively of the differences between the measurements of the two readers. The value of δ is computed as a factor multiplied by the mean of the more experienced reader's

measurements. In this study, the factor was 0.20, 0.25, or 0.30. The variable δ accounts for the expected variability in the measurements made by the two radiologists. A smaller value of δ implies stricter criteria for demonstrating that the measurements of the two radiologists are equivalent.

The *ICC* coefficient is also used to report the degree of agreement between multiple readers. A number of different models can be used for computing the *ICC* value [64]. In this study, to report the inter-observer agreement, a two-way random model was used since the set of images is a random subset of images from the class of mammographic images and the radiologists are also randomly selected from the population of radiologists. The *ICC* coefficient is defined as follows:

$$ICC = \frac{MS_E - MS_R}{MS_R + (k - 1) \cdot MS_E + \frac{k}{n} \cdot (MS_C - MS_E)} \quad (2.2)$$

Where k denotes the number of readers, n denotes the number of images, MS_R is the mean square error between images, MS_E is the residual mean square error, and MS_C is the mean square error between readers. For the computation of the intra-observer agreement, a two-way mixed model was used as the set of images is considered as a random subset of images from all mammographic images but the measurements are made by a single radiologist and thus the rater (radiologist) is considered as a fixed effect in the *ICC* model [64].

Different guidelines exist for the interpretation of *ICC*, but one reasonable scale is that an *ICC* value of less than 0.40 indicates poor reproducibility, *ICC* values in the range 0.40 to 0.75 indicate fair to good reproducibility, and

an *ICC* value of greater than 0.75 shows excellent reproducibility [68].

Bland-Altman analysis (also known as the method of differences) has been proposed for measuring the degree of agreement [65, 66]. In this method, the differences in the measurements made by two readers are plotted against the average values of these measurements. According to Bland and Altman [65, 66], if 95% of the differences are within ± 1.96 standard deviations of the mean of the differences, then this denotes good agreement between the two sets of measurements. These limits are also known as the “limits of agreement”. Note that hypothesis testing for equivalence and the *ICC* method provide quantitative measures of the agreement between the measurements whereas the Bland-Altman analysis technique provides a qualitative assessment.

2.3 Results

The inter- and intra-observer variability of measurements of spiculated masses was evaluated using a hypothesis test for equivalence, the *ICC* coefficient, and the Bland-Altman technique. For an equivalence level of 30% of the mean of R1’s first measurement (Table 2.2), equivalence was achieved between R1’s and R2’s measurements ($N = 21$) for average spicule length ($p < 0.01$), average spicule width ($p = 0.03$), and the count of the number of spicules ($p < 0.01$). For comparing the major axis measurements, one case was removed since R2 inadvertently measured the minor axis; with $N = 20$, equivalence was achieved for the length of the major axis ($p < 0.01$).

Similarly, Table 2.3 shows the degree of agreement between the mea-

Total no. of cases	Delta	Major Axis	Spicule Width	Spicule Length	No. of Spicules
		R1's mean = 3.78	R1's mean = 0.278	R1's mean = 2.44	R1's mean = 17.57
		R2's mean = 3.73	R2's mean = 0.221	R2's mean = 2.39	R2's mean = 18.48
21 \$	$\delta = 0.30 \cdot \text{mean of R1's measurement}$	p < 0.01	p = 0.03	p < 0.01	p < 0.01
21 \$	$\delta = 0.25 \cdot \text{mean of R1's measurement}$	p < 0.01	p = 0.18	p < 0.01	p < 0.01
21 \$	$\delta = 0.20 \cdot \text{mean of R1's measurement}$	p < 0.01	p = 0.54	p < 0.01	p < 0.01

Table 2.2: This table provides the summary statistics and the results of the hypothesis test for equivalence between the measurements of radiologists R1 and R2. The null hypothesis was that the two radiologists are not equivalent. Thus, if we obtain a p-value of less than 0.05 (bold type), we can reject the null hypothesis and say that the measurements of the two radiologists are equivalent. (\$ One observer measured minor axis by mistake, so that image was removed for the major axis calculation only.)

Total no. of cases	Major Axis	Spicule Width	Spicule Length	No. of Spicules
21\$	$ICC = 0.801$	$ICC = 0.561$	$ICC = 0.770$	$ICC = 0.780$

Table 2.3: (Inter-Observer agreement) Intra-class correlation coefficients for the measurements made by radiologist R1 and R2. (\$ One observer measured minor axis by mistake, so that image was removed for the major axis calculation only.)

measurements of the radiologists R1 and R2 using the ICC method. Our analysis shows that there is “excellent” inter-rater agreement between R1’s and R2’s measurements ($N = 21$) for average spicule length ($ICC = 0.770$), and the count of the number of spicules ($ICC = 0.780$). “Fair to good agreement” was obtained for the average spicule width ($ICC = 0.561$). For comparing the major axis measurements, one case was removed and with $N = 20$. “Excellent” inter-rater agreement was observed for the length of the major axis ($ICC = 0.801$). The inter-observer agreement was also analyzed using the Bland-Altman technique. Bland and Altman suggested that if 95% of the differences were within the “limits of agreement” then this denoted good agreement between the two sets of measurements. According to the Bland-Altman method, good inter-observer agreement is obtained for all four parameters measures (Figure 2.3).

We studied the intra-observer variability based on re-measurement of 12 images by the senior radiologist R1. For an equivalence level of 30%, equivalence was achieved between R1’s first and second measurements ($N = 12$) for all properties (Table 2.4): average spicule length ($p < 0.01$), average spicule

Total no. of cases	Delta	Major Axis	Spicule Width	Spicule Length	No. of Spicules
12	$\delta = 0.30 \cdot \text{mean of R1's measurement}$	p = 0.00	p = 0.00	p = 0.00	p = 0.00
12	$\delta = 0.25 \cdot \text{mean of R1's measurement}$	p = 0.00	p = 0.00	p = 0.02	p = 0.02
12	$\delta = 0.20 \cdot \text{mean of R1's measurement}$	p = 0.01	p = 0.00	p = 0.26	p = 0.09

Table 2.4: Results of the hypothesis test for equivalence between the first and second set of measurements made by radiologists R1. The null hypothesis was that the two sets of the measurements made by radiologist R1 are not equivalent. Thus, if we obtain a p-value of less than 0.05 (bold type), we can reject the null hypothesis and say that the two sets of measurements are equivalent.

width ($p < 0.01$), length of major axis ($p < 0.01$), and the count of the number of spicules ($p = 0.01$). Moreover, equivalence was demonstrated even at the stricter level of 25% of the mean of R1's first measurement.

The intra-observer agreement between the two sets of measurements made by radiologist R1 ($N = 12$) using the *ICC* method were also very good (Table 2.5). The intra-observer agreement was “excellent” for the length of the major axis ($ICC = 0.951$), average spicule length ($ICC = 0.852$), and the average spicule width ($ICC = 0.896$). “Fair to good agreement” was observed for the count of the number of spicules ($ICC = 0.641$). The intra-observer agreement was also analyzed using the Bland-Altman technique. Bland and Altman suggested that if 95% of the differences were within the “limits of agreement” then this denoted good agreement between the two sets of measurements. Figure 2.4 show that for all of four parameters measured, good intra-observer

Total no. of cases	Major Axis	Spicule Width	Spicule Length	No. of Spicules
12\$	$ICC = 0.951$	$ICC = 0.896$	$ICC = 0.852$	$ICC = 0.641$

Table 2.5: (Intra-Observer agreement) Intra-class correlation coefficients for the two sets of measurements made by radiologist R1

agreement is obtained according to the Bland-Altman technique.

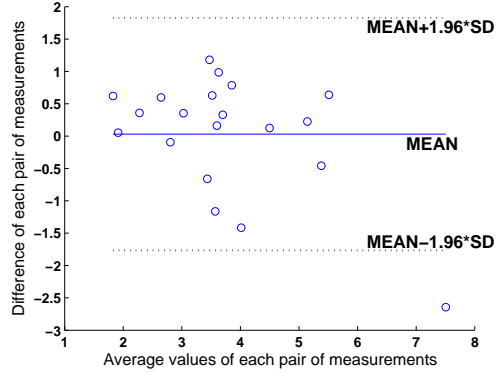
2.4 Discussion

In this chapter, it was shown that it is feasible to make reliable measurements of the physical properties of spiculated masses on mammography. The properties measured were the length and width of all spicules and length of the major axis of the central mass region. The count of the total number of spicules was also assessed. In this study, we obtained good inter- and intra-observer agreement for the measurement of these properties of spiculated masses. This was demonstrated with a hypothesis test for equivalence, the intraclass correlation coefficient, and the Bland-Altman analysis. Since such a measurement task is not a part of the radiologists' regular clinical duties, the training stage was crucial for this measurement study. In the training phase, the radiologists discussed measurements that they had made independently (Figure 2.2(a)). While it was difficult for them to verbalize a consensus measurement protocol, the discussion was clearly fruitful since the data collected for the training process did not show equivalence (except for major axis), but equivalence was demonstrated for all four physical parameters in the measurement study after the training was complete.

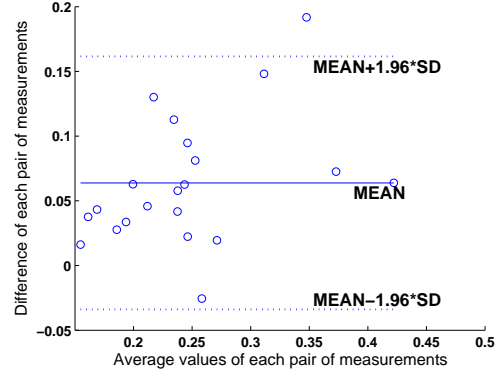
Two interesting points are evident from a visual inspection of the marked images from the training and measurement phases of the study. First, some of the changes to their measurement protocol can be surmised; before the training, R2 typically marked spicules as being much longer than R1, but R2 marked the spicule lengths similarly to R1 after the training phase. Second, we noticed was that if the two readers picked the same spicule, their measurements for that spicule were nearly identical. Thus, the primary source of variability appears to be the identification of structures as “spicules” rather than the task of measuring a spicule after it is located. Both of these points are observed in Figure 2.2, where the measurements made by the two radiologists are overlaid on the original image. Figure 2.2(a) shows the measurements made on an image during the training stage and (Figure 2.2(b)) shows the measurements on an image from the second set of spiculated masses.

To the best of our knowledge, no prior study has measured the physical properties of masses on mammograms or assessed the observer variability of such a task. However, researchers have reported the inter- and intra-observer agreement for various measurements tasks in other areas of medical imaging (e.g., [55–61]). Though several statistical methods can be used to report the inter-observer agreement, the most common approach has been to use the *ICC*. The *ICC* values reported in prior medical imaging measurement studies range from 0.570 to 0.820; thus, the *ICC* values observed this study (0.561 to 0.951) are within the range defined by previous work. Thus, we have demonstrated that properties of spiculated masses can be reliably measured on

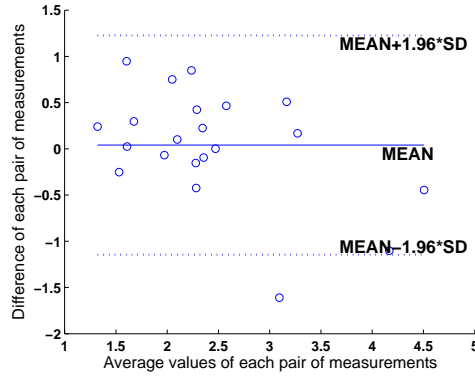
mammography, within the level of inter- and intra-observer variability typical of other measurement tasks in radiology.



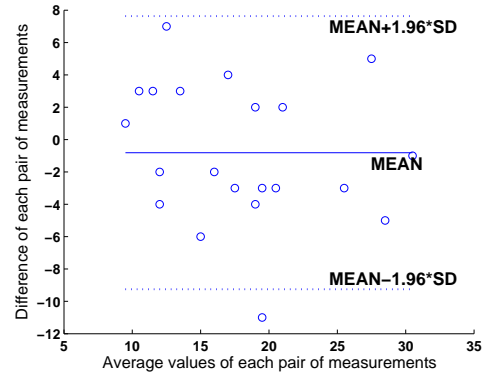
(a) Bland-Altman analysis for the major axis of the spiculated masses



(b) Bland-Altman analysis for the spicule width



(c) Bland-Altman analysis for the spicule length



(d) Bland-Altman analysis for the number of spicules of the spiculated masses

Figure 2.3: Bland-Altman analysis for the inter-observer agreement for each of the four physical characteristics that were measured by radiologists R1 and R2. The parameters measured were: (a) major axis of the spiculated masses, (b) the width of the spiculations, (c) the length of the spiculations, and (d) the number of spiculations.

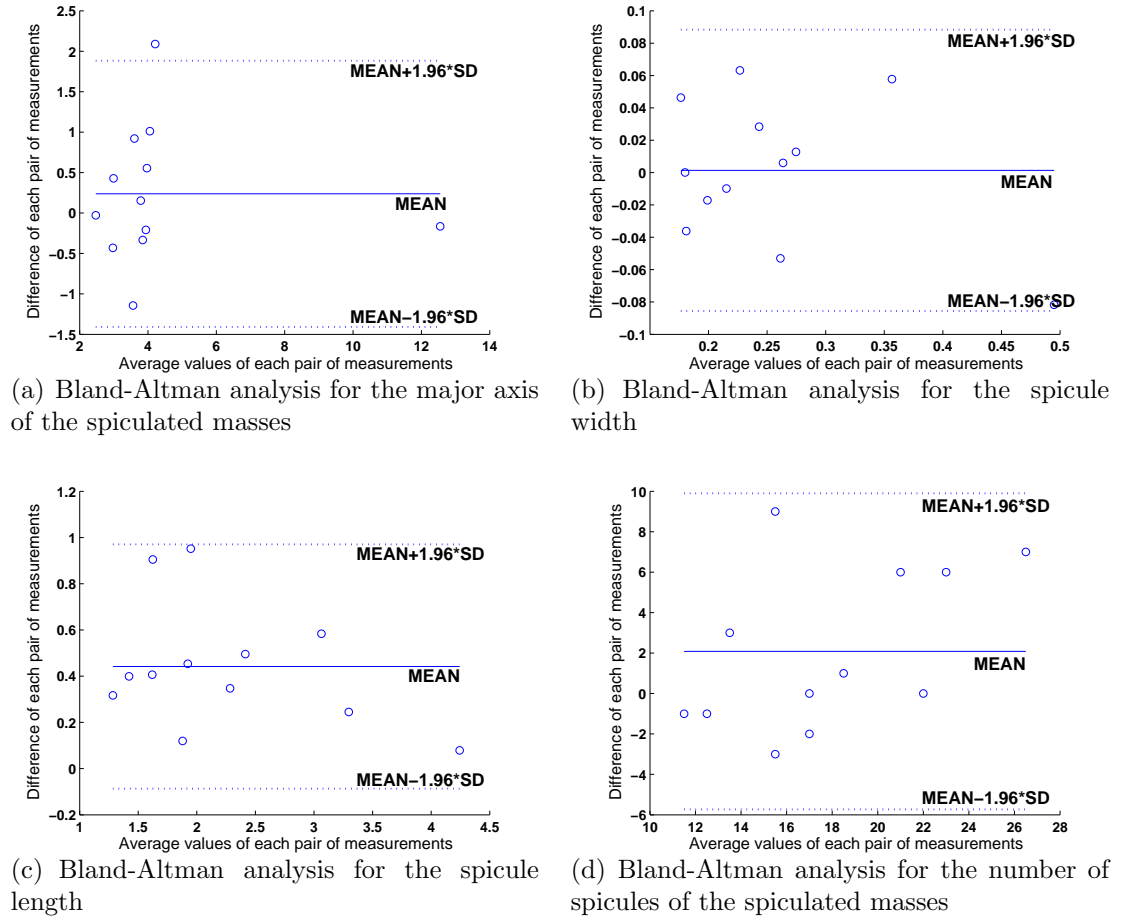


Figure 2.4: Bland-Altman analysis for the intra-observer agreement for each of the four physical characteristics that were measured twice by the senior radiologist R1. The parameters measured were: (a) major axis of the spiculated masses, (b) the width of the spiculations, (c) the length of the spiculations, and (d) the number of spiculations.

Chapter 3

Detection Algorithm

3.1 Introduction

This chapter presents the detection algorithm that was developed during the course of this dissertation. The detection algorithm consists of three main components, which are:

1. Spiculation filtering to detect the spatial locations where spicules converge.
2. Detection of the central mass region of the spiculated masses.
3. Suppressing false-positive due to normal linear structures.

Each of these components targets a specific aspect of spiculated lesions. The first component, Spiculation filtering is aimed towards determining the spatial locations where spicules converge. The second component uses Gaussian filters to detect the central mass regions of spiculated masses. It was observed that normal linear structures such as blood vessels, ducts and connective tissue can cause false-positives. The last component was geared toward suppressing false-positives due to these linear structures. Figure 3.1 shows the block diagram of

the proposed detection algorithm and each of the components of the detection algorithm are described in detail in sections 3.2, 3.3 and 3.4 respectively.

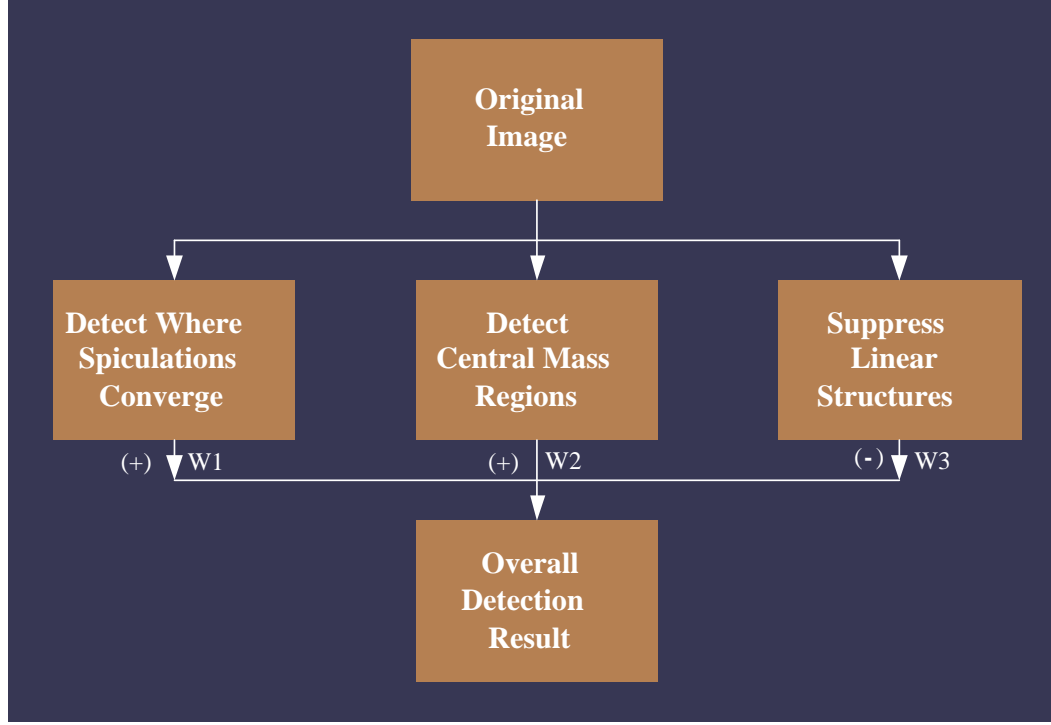


Figure 3.1: This figure shows the block diagram of the detection algorithm. This algorithm consists of three components. The first component aims to detect the spatial locations where spicules converge. The second component is designed to detect the central mass regions of the spiculated masses and the third component aims to reduce false-positives due to normal linear structures.

3.2 Overview of Spiculation Filtering

This section presents the details of the spiculation filtering algorithm which aims to find the spatial locations where spicules converge. The spiculation filtering algorithm for the detection of spiculated lesions is based on

measurements of their physical characteristics and consists of two steps: an enhancement step followed by a filtering/detection step. The enhancement step uses a new technique for the enhancement of spiculations in which a linear filter is applied to the Radon transform of the image. In the second step, the enhanced images are filtered with a new class of linear image filters, which we call Spiculation Filters. These filters have been invented specifically for detecting spiculated lesions, which are characterized by converging lines or spiculations. The filter parameters for both steps were determined based on measurements made on a set of spiculated masses and architectural distortions.

3.2.1 Enhancement of Spiculations

The most prominent feature of spiculated masses and architectural distortions is the presence of spicules radiating in all directions. The goal of this step is to enhance spicules which are approximated as linear structures. Specifically, we model spicules as lines of widths that fall within a predetermined range determined by prior physical measurements on real spicules.

The Radon domain is a convenient space in which to detect lines. To enhance spicules, we have developed a novel algorithm in which we compute the Radon transform of the image and then perform filtering in the Radon domain to enhance linear structures that fall within a specified, evidence-based range of widths.

The Radon transform of a continuous function $f(x, y)$ is defined as [69] :

$$g(\rho, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \delta(\rho - x \cos(\theta) - y \sin(\theta)) dx dy \quad (3.1)$$

where $\delta(r)$ is the Dirac delta function [70]. In equation (3.1), ρ and θ are parameters of the Radon domain. As can be seen from Fig. 3.2, for any given line in an image, ρ is the length of a perpendicular line segment from the origin to the line and θ is the orientation of the line segment with respect to the x -axis. The corresponding discrete Radon transform for an N -by- N image is:

$$\hat{g}(\rho, \theta) = \sum_{y=-N/2}^{N/2} \sum_{x=-N/2}^{N/2} f(x, y) \delta(\rho - x \cos(\theta) - y \sin(\theta)) \quad (3.2)$$

where $\delta(r)$ is the Kronecker delta function [70]. The center of the image is at $(0, 0)$ and the x and y co-ordinates extend from $-N/2$ to $N/2$. The parameter θ was varied from 0 to 179 degrees and the parameter ρ was varied from $-N/\sqrt{2}$ to $N/\sqrt{2}$. The Hough transform is closely related to the discrete Radon transform. The Hough transform can be viewed as the discrete Radon transform of a binary image [71]. The term $\delta[\rho - x \cos(\theta) - y \sin(\theta)]$ in (3.1) computes the integral of $f(x, y)$ along the line $(\rho = x \cos(\theta) + y \sin(\theta))$ and thus the value of $g(\rho, \theta)$, for any (ρ, θ) is the integrated density of $f(x, y)$ along this line. Thus, a line in the image space $f(x, y)$ produces or maps to a point in the Radon domain.

Also note that lines of different thickness will have different representations in the Radon domain. Figures 3.2(a) and 3.2(b) show two lines of

different thickness and their corresponding Radon transforms. In Fig. 3.2(a), the line is 1 pixel thick, and in Fig. 3.2(b), the line is 4 pixels thick. For simplicity, it is assumed that the two lines have the same length of 255 pixels and that each pixel on the line has a uniform intensity of 1. In the Radon domain, the row values correspond to the different values of ρ and the column values represent the different θ values. Thus, a single pixel thick line would be represented by a point in the Radon domain, whereas a 4 pixel thick line would be represented by 4 points along a column in the Radon domain. Thus by enhancing sharp changes or “local peaks” along the columns in the Radon domain, it is possible to detect the corresponding lines in the image.

To do so, a peak detection algorithm needs to be applied. While many approaches might be taken to identify peaks, one approach is to simply filter the columns of the Radon transform with an appropriate bi-valued one-dimensional linear filter with an impulse response that is a rectangle of width agreeing with the maximum spicule width to be highlighted. This approach has the advantages of both directness and simplicity. Application of this simple filter to the Radon domain, followed by an inverse Radon transform, will yield an enhanced image with amplified linear structures (spicules) of the requisite widths, with all other structures being suppressed. The particular filter was chosen based on the thickness of the spicules obtained from our preliminary measurement studies. More details on the selection of this filter are given in Section 3.5.2.

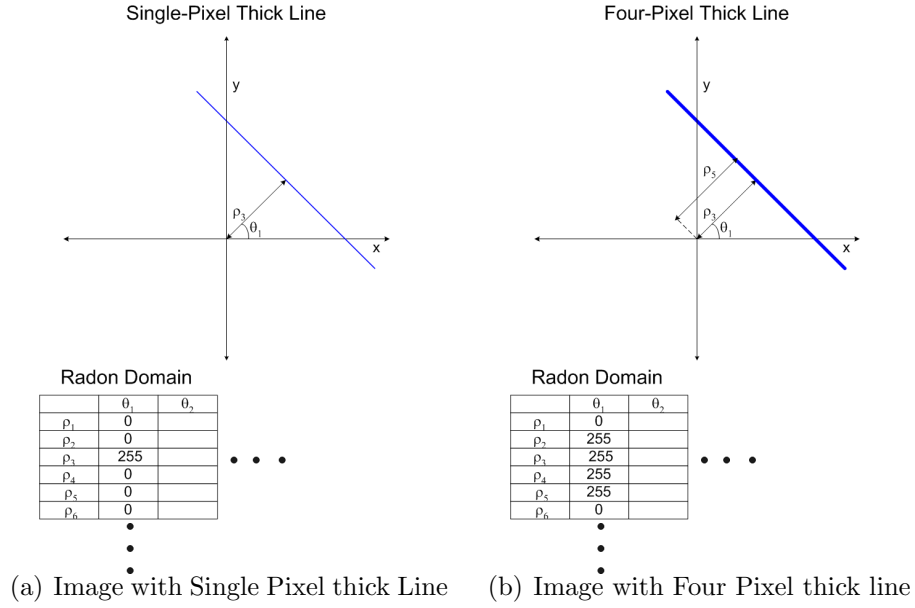


Figure 3.2: These images show lines of different widths and the corresponding section of their Radon transforms. For simplicity, we assume that each of these lines is 255 pixels long and the intensity at each pixel is one. Thus, peaks of value 255 occur in the Radon domain. A key point is that lines of different widths have different representations in the Radon domain.

For our application, the filter: $[-1 \ -1 \ 1 \ 1 \ 1 \ 1 \ 1 \ -1 \ -1 \ -1]^T$, was used to detect peaks along the columns in the Radon domain. This is achieved by convolving each column with this filter. Note that the coefficients at the center of the filter are positive, while those at the periphery are negative and the sum of coefficients of the filter is zero. Thus, the response of this filter to an area of constant or slowly varying values will be zero or very small. After filtering in the Radon domain, the inverse Radon transform is computed using the filtered back-projection algorithm [72] to obtain an image in which linear

structures have been enhanced:

$$f(x, y) = \int_0^\pi Q_\theta(x \cdot \cos(\theta) + y \cdot \sin(\theta)) d\theta \quad (3.3)$$

$$Q_\theta(t) = \int_{-\infty}^{\infty} S_\theta(w) |w| e^{j \cdot 2\pi w t} dw \quad (3.4)$$

The enhanced image $f(x, y)$ is obtained using (3.3), where $Q_\theta(t)$ is defined in (3.4) and $S_\theta(w)$ is the 1-D Fourier transform of $g_\theta(r)$ [72]. The corresponding equations for the inverse discrete Radon transform (DRT) are:

$$\hat{f}(x, y) = \frac{\pi}{K} \sum_{i=1}^K \hat{Q}_{\theta_i}(x \cdot \cos(\theta_i) + y \cdot \sin(\theta_i)) \quad (3.5)$$

$$\hat{Q}_\theta \left(\frac{k}{2W} \right) \approx \frac{2W}{N} \sum_{m=-N/2}^{N/2} S_\theta \left(m \cdot \frac{2W}{N} \right) \left| m \frac{2W}{N} \right| e^{j 2\pi (mk/N)} \quad (3.6)$$

Where $\hat{f}(x, y)$ is the image obtained by computing the inverse DRT. It is assumed that $S_\theta(w)$ is bandlimited and is zero outside the interval $(-W, W)$ for all values of θ . Thus, the output of the enhancement stage is a line-enhanced image and all subsequent processing is performed on this image. It is known that reconstructing images from a few projections causes artifacts [73]. To avoid reconstruction artifacts, the parameter θ was varied from 0 to 179 degrees in increments of 0.1 degrees during the computation of the Radon transform.

The Radon transform and the Hough transform have been applied for the detection of spiculated masses by other researchers [35]. However, the key engineering innovation of our enhancement algorithm is the filtering of

the coefficients in the Radon domain. To the best of our knowledge, this has not been proposed in CADe or other image processing applications. This allows the enhancement of lines of different widths by choosing the appropriate parameters of the filter. Figure 3.3 shows the various steps of the enhancement algorithm.

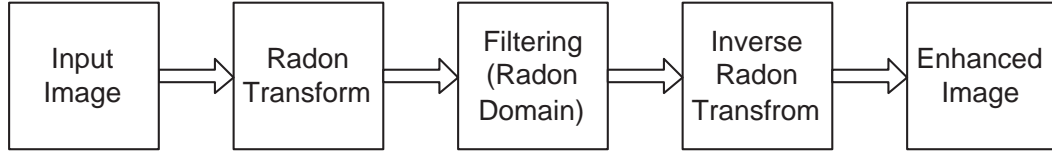


Figure 3.3: Flow-chart of the Enhancement Algorithm: First, the Radon transform of an image is computed. The Radon domain is then filtered with a column filter. Finally, the enhanced image is obtained by computing the inverse Radon transform.

3.2.2 Detection Using Spiculation Filters

The goal of the enhancement stage described in Section 3.2.1 was to enhance spiculations on mammograms. The aim of this stage is to detect the spatial locations where these spiculations converge. For this purpose, we have invented a new class of wavelet-type filter banks that are specifically tuned to match the physical structures that define spiculated lesions. These new filters, which we have recently conceived [53, 74] are quite unique in that they are designed to match the physical structure of spiculated lesions. These filters were parameterized using data collected from our preliminary study on the quantitative morphology of spiculated lesions.

The fundamental idea is to design a “matched filter” for a spiculated lesion. That is, once the enhanced image is processed with these filters, a large output would be obtained when the filters are “tuned to” or “match” a spiculated lesion. The filters are a new class of complex quadrature Spiculation Filters (SFs) that are specifically responsive to spicule groupings, having a distribution of orientations and positions such that they converge to a central region or radiate from that region. The complex SF has the advantage that the sum of the squared responses of the two quadrature components is phase independent. Thus, the filter will have the same response to a spiculation even if it has been rotated.

A Spiculation Filter (SF) has a filter impulse response with an appearance entirely different from anything previously conceived of: a Gaussian-modulated sine torus. The SF consists of quadrature components which are the so-called cosine SF ($f_{\cosine}(r, \theta; r_0, \sigma, \omega)$) and the sine SF ($f_{\sine}(r, \theta; r_0, \sigma, \omega)$). These components are shown in Fig. 3.4 and are defined as follows:

$$f_{\cosine}(r, \theta; r_0, \sigma, \omega) = g(r; r_0, \sigma) * \cos(\omega\theta) \quad (3.7)$$

$$f_{\sine}(r, \theta; r_0, \sigma, \omega) = g(r; r_0, \sigma) * \sin(\omega\theta) \quad (3.8)$$

$$g(r; r_0, \sigma) = \exp[-(r - r_0)^2 / (2 * \sigma^2)] \quad (3.9)$$

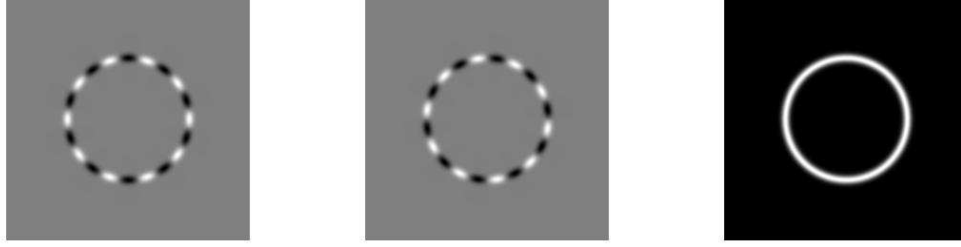
where: $r = \sqrt{x^2 + y^2}$, $\theta = \arctan(y/x)$ and σ is the standard deviation in pixels. The parameter r_0 is a size parameter measured in pixels, and ω is the modulation frequency measured in cycles per circumference. The Gaussian torus is defined by equation (3.4(c)). The SFs can also be written in the

overall phasor form as follows:

$$h(r, \theta; r_0, \sigma, \omega) = f_{\cosine}(r, \theta; r_0, \sigma, \omega) + j * f_{\sine}(r, \theta; r_0, \sigma, \omega) \quad (3.10)$$

Also note that the RMS value $\sqrt{(f_{\cosine}^2 + f_{\sine}^2)}$ of the quadrature components is equal to the toroidal Gaussian envelope function $g(r; r_0, \sigma)$.

The toroidal Gaussian in Fig. (3.4(c)), is responsive to structures that intersect a circular spatial band at any point in the image. A perfect “match” to such a filter would be a similar band occurring anywhere in the image. The key modification to the filter is the sinusoidal modulation, which creates the alternating bands in the Gaussian. The period or frequency of the bands is controlled by the parameter ω . The physical significance of the alternations is that each filter now becomes responsive to oscillatory structures occurring along circles in the image. Such a filter will be responsive to a band passing through a spiculated lesion candidate. Ideally, we would like the SF to match the spiculated lesions exactly, but this is not possible since real spiculated lesions present highly random arrangements of spicule spacings and spicule widths. Nevertheless, the SF do possess important optimality properties: since they are Gaussian modulated sinusoids, they have excellent spatial localization properties, as well as excellent spectral localization among filters that can be written in the form $h(r, \theta; r_0, \sigma, \omega) = f_{\cosine}(r, \theta; r_0, \sigma, \omega) + j * f_{\sine}(r, \theta; r_0, \sigma, \omega)$. They are, in a sense, radial Gabor functions, which optimize the space-frequency uncertainty tradeoff in achieving maximum conjoint resolution; because of this, Gabor functions have proven to be exceedingly effective for image analysis



(a) cosine SF with $\sigma = 2, f = 10$ (b) sine SF with $\sigma = 2, f = 10$ (c) Gaussian torus $\sigma = 2$

Figure 3.4: Example of a Spiculation Filter (SF): Figures 3.4(a) and 3.4(b) show the two quadrature components of the SF, namely, the cosine SF and the sine SF, respectively. The toroidal Gaussian envelope of the SF is shown in Fig. 3.4(c).

applications [75–78].

Like Gabor filters (as applied to “standard” image textures), SFs can be expected to extract relevant spiculation features (quantified as SF responses) with the maximum possible spatial resolution, for a given filter bandwidth (frequency domain coverage). Thus, not only are SFs optimally matched to the expected shapes of spiculations (as defined by our model), but they will also detect these shapes with optimal resolution (positional accuracy and detail sensitivity).

The response of a SF to a spiculated lesion is important, but it supplies incomplete information regarding the pattern. Instead, only a band of a given radius is analyzed. However, by combining multiple SF in a filter bank, overall spiculated lesion responses may be obtained. Thus, the next innovation that

we have developed is the concept of a Spiculation Filter bank (SFB). Two such SFBs are shown in Fig. 3.5.

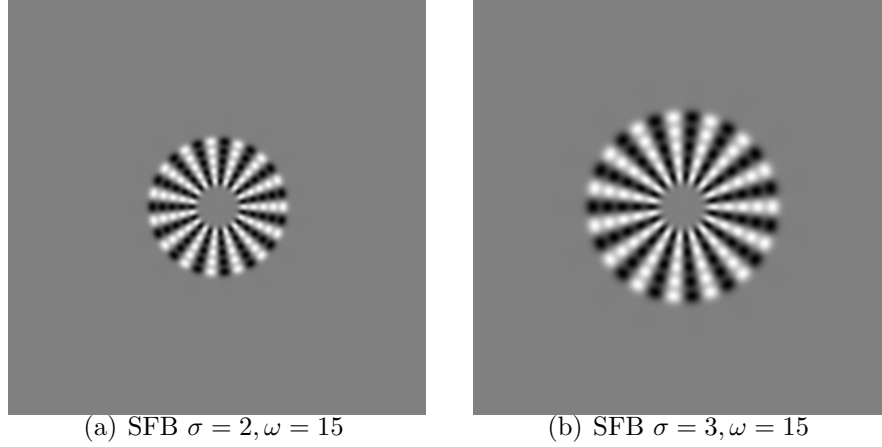


Figure 3.5: Spiculation FilterBanks (SFBs): Two Spiculation Filter banks, each composed of five SF of progressively increasing radii and matching radial frequencies, ω . The SFB in Figs. 3.5(a) and 3.5(b) have the same inner radius but different outer radii. These would correspond to lesions where the central mass is the same size but the spicule lengths differ.

Here, the constituent SF have progressively increasing radii but the same frequencies. In such a filter bank, the radius of the central region, whether a radio-opaque mass or not, can be made to match the SFB by appropriate selection of the inner filter radius. Likewise, spicule length can be matched to the SFB by appropriate selection of the inner and the outer filter radii, the difference between these corresponding to spicule length. Of course, the physical parameters of a particular lesion cannot be known beforehand, and hence a single appropriate SFB cannot be predetermined. However, we use the measurement data (Chapter 2), taken from a large, ongoing study in-

volving multiple radiologists, to support the selection of the inner and outer radii r_{inner} and r_{outer} for each SFB, as well as the frequency ω of each SFB. We will then be supplied with a collection of SFB which can be applied to each mammographic image with a high certainty that the SFB being applied will match the physical characteristics of any existing spiculated lesion in the breast image.

In the SFB pictured in Fig. 3.5, the filters were chosen such that the component toroidal Gaussian envelopes intersect each other at their half-peak values, a strategy that ensures that there will be no gaps in the responses along the radial directions. This approach has been shown to be efficacious in numerous studies involving Gabor filter applications [75, 76].

Thus, the relationship between the radii of two adjacent SF of a SFB is given by the following: Let the radius of a single SF be r_1 and the toroidal component of that filter is given by $g(r; r_1, \sigma)$. Let r_2 be the radius of the next SF in the filter bank. Then r_2 can be computed by the following equation:

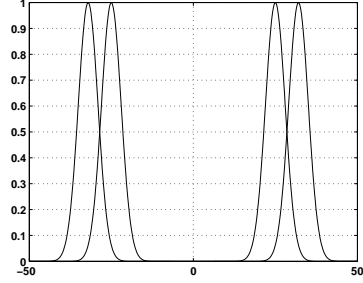
$$r_2 = r_1 + 2.3548 * \sigma \quad (3.11)$$

Figure 3.6(a) shows the toroidal components of two SFs of a filter bank and Fig. 3.6(b) shows a cross-section through these components.

We now describe how a SFB can be used for the detection of spiculated lesions. First, the enhanced image is filtered with each SFB. The normalized cross correlation (NCC) was used for filtering the enhanced image with each SFB. Let T be a SFB of size $M \times M$ and let I be the image of size $L \times L$ such



(a) The Gaussian toroidal components of two filters



(b) A cross sectional view through those two components

Figure 3.6: Figure 3.6(a) shows the toroidal components of two SFs of a filter-bank and Fig. 3.6(b) shows a cross-section through these components. The filters are chosen such that the component toroidal Gaussian envelopes intersect each other at their half-peak values, a strategy that ensures that there will be no gaps in the responses along the radial directions.

that $M \ll L$. The NCC is defined as follows:

$$NCC(i, j) = \frac{1}{\sigma_T \cdot \sigma_I} \sum_{k=1}^M \sum_{m=1}^M \{(T(k, m) - \mu_T) (I(i + k - 1, j + m - 1) - \mu_I(i, j))\} \quad (3.12)$$

$$\sigma_T = \sqrt{\sum_{k=1}^M \sum_{m=1}^M (T(k, m) - \mu_T)^2}$$

$$\sigma_I = \sqrt{\sum_{k=1}^M \sum_{m=1}^M (I(i + k - 1, j + m - 1) - \mu_I(i, j))^2}$$

Where μ_T is the mean of the template and $\mu_I(i, j)$ is the mean of the sub-image centered at (i, j) .

Let OP_{sin} and OP_{cos} be the outputs obtained after filtering the enhanced image with the two quadrature components of the SFB (cosine and

sine SFBs). The magnitude response OP_{mag} and the phase response OP_{phase} of the SFB are defined as:

$$OP_{mag} = \sqrt{OP_{cos}^2 + OP_{sin}^2} \quad (3.13)$$

$$OP_{phase} = \arctan(OP_{sin}/OP_{cos}) \quad (3.14)$$

If the filter bank “matches” a spiculated lesion, a peak would be obtained at the spatial location corresponding to the center of the spiculated lesion. Thus, suspicious regions can be identified by detecting the local peaks in the overall output. Note that the larger the overall output at a particular spatial location, the higher the likelihood that the spatial location corresponds to the center of a spiculated lesion.

The procedure described above is repeated for all filter-banks and these are combined to form a new image (*SFB_Maxima*) in the following way. Let $SFB1, SFB2, \dots, SFBN$ denote the N SFBs. In the *SFB_Maxima* image the value at each spatial location corresponds to the maximum filter-bank output (across all SFBs).

$$SFB_Maxima(x, y) = \max \{SFB1(x, y), SFB2(x, y), \dots, SFBN(x, y)\} (\forall x, y) \quad (3.15)$$

In the *SFB_Maxima* image, the first N marks are selected as the possible locations of suspicious masses. The aim in this step is to detect as many suspicious regions as possible in order to achieve a high detection sensitivity. This may lead to a large number of FPI. This is acceptable, since additional steps will subsequently be taken to reduce the number of FPI.

3.3 Detecting central mass regions

It is known that all spiculated masses have a central mass region and thus the second component of the detection algorithm is geared towards the detection of central mass regions. To detect these regions we used Gaussian filters. Gaussian filters have been previously used for detection of masses. However, the key differences in our approach are that the size of these filters were selected from the measurement data and that a bank of Gaussian filters were used.

The proposed central mass region detection algorithm was as follows: Each image was filtered using a bank of 2-D Gaussian filters. The sigmas are chosen such that the half-peak radii (HPR) of the Gaussian filters are 10, $10\cdot\sqrt{2}$, 20, and $20\cdot\sqrt{2}$ and 40 pixels respectively. The HPR was the distance where the value of the Gaussian filter falls to half it's maximum value. These radii are reflective of the measurement data. As the half peak value occurs at $1.1774 \cdot \sigma$, each sigma value was computed as follows : $\sigma = HPR/1.1774$. The filter-size was set to be $6 \cdot \sigma$. These filters are shown in figure 3.7. The original image is filtered with each of these Gaussian filters and at each pixel, the maximum output value across all Gaussian filters is computed. The pixels with a high output represent presence of mass. Let this output be image C . As the coefficients in image C indicate the presence of a mass, these are summed with the coefficients from the filter-bank outputs. Thus the combined output (Op) is now defined as: ($Op = A - B + C$).

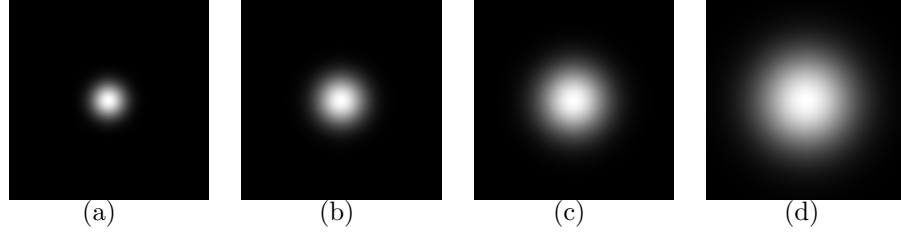


Figure 3.7: This figure shows the set of 4 Gaussian filters used. The half-peak radius of these filters are $10 \cdot \sqrt{2}$, 20 , $20 \cdot \sqrt{2}$, and 40 pixels respectively.

3.4 Suppression of linear structures

In the spiculation filtering component described in section 3.2, the focus was on the enhancement of spicules, followed by locating the spatial locations where the spicules converge. In addition to the spicules, the enhancement technique also enhances linear structures such as blood vessels, ducts other other linear structures in the tissue parenchyma. The Spiculations FBs responds to the criss-crossings of these structures and thus lead to false-positives. To remove the false-positives due to these linear structures, we used oriented difference-of-Gaussian filters to identify the locations of these linear structures and then suppressed the false-positives at these locations using the following strategy.

At each spatial location (i, j) in the image, the local orientation $\theta_{i,j}$ was computed as:

$$\theta_{i,j} = \arctan(Gy/Gx) \quad (3.16)$$

Where Gx and Gy are the gradients in the x and y directions respectively.

At each spatial location the image was filtered with an elongated and oriented Difference-of-Gaussian (DoG) filter. The orientation of this filter was the local orientation at that spatial location as computed from equation 3.16. A subset of such oriented DoGs are shown in Figure 3.8. The oriented DoGs were created as the difference of two oriented Gaussian filters. These oriented-DoG filters are also evidence-based since the width of the central lobe is greater than the largest width of the spicules that was measured. In addition, their length is greater than the maximum spicule length that was measured. (As the goal is to suppress linear structures that are not spicules.) This gives us an image in which curvi-linear structures such as ducts and blood-vessels are enhanced. Let this image be B . Figure 3.9 shows examples of images where these linear structures have been enhanced. As we want to suppress any FP's that occur at these locations, this output is combined with the output of the SFBs in the following way. Let A be the image of the maximum of the filter-bank outputs at each spatial (obtained by finding the maximum at each spatial location). The combined output (Op) is now defined as: ($Op = A - B$).

3.5 Methods

3.5.1 Data Description

The images for testing the detection algorithm were obtained from the Digital Database for Screening Mammography (DDSM) [62]. The DDSM is the largest publicly available dataset of digitized mammograms. The entire database consists of 2,620 cases. Each case consists of four mammograms: a

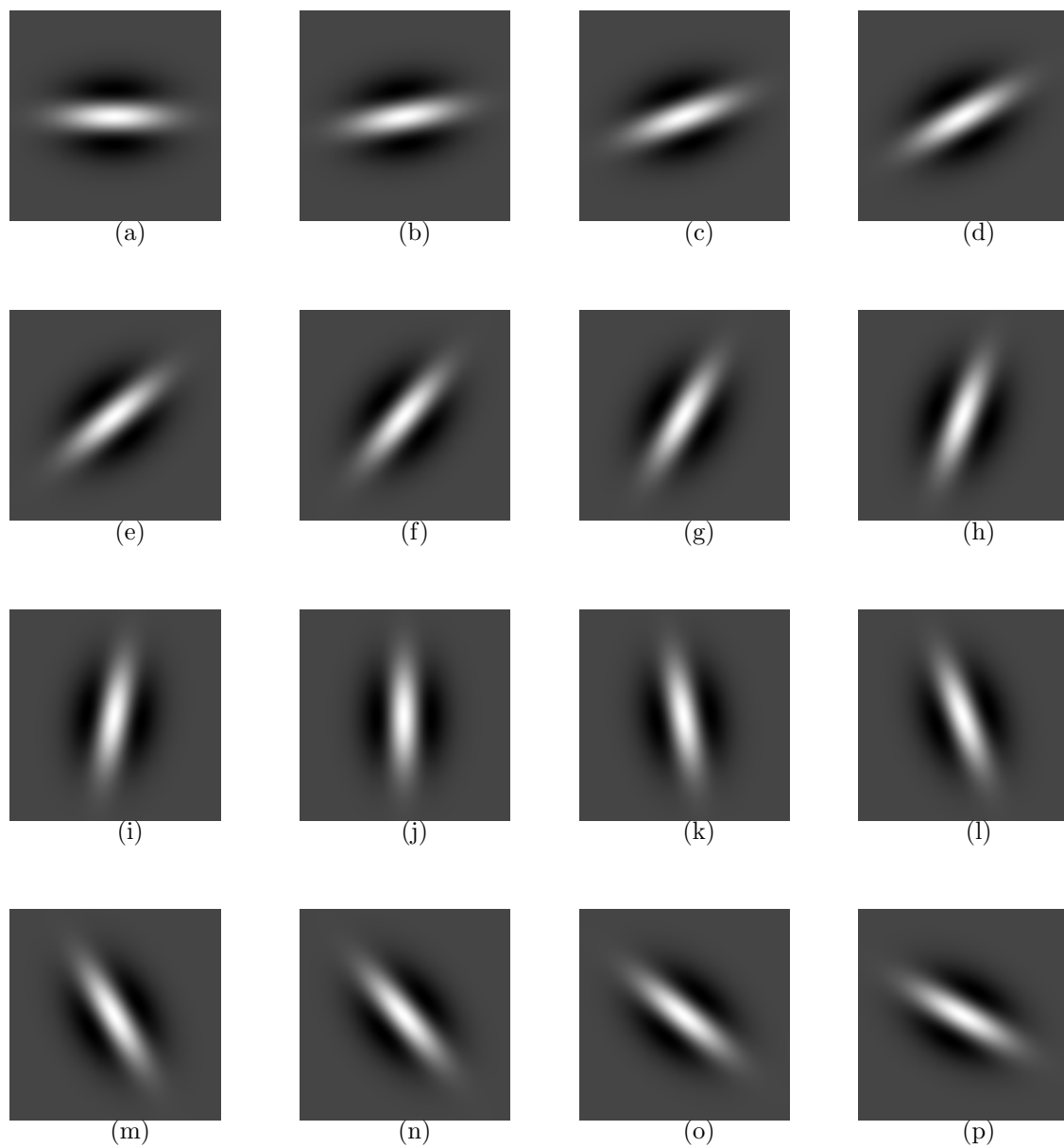


Figure 3.8: This figure shows a subset of the elongated and oriented DoG filters used. The output of these filters was used to suppress the FPs due to linear structures such as blood vessels and ducts.

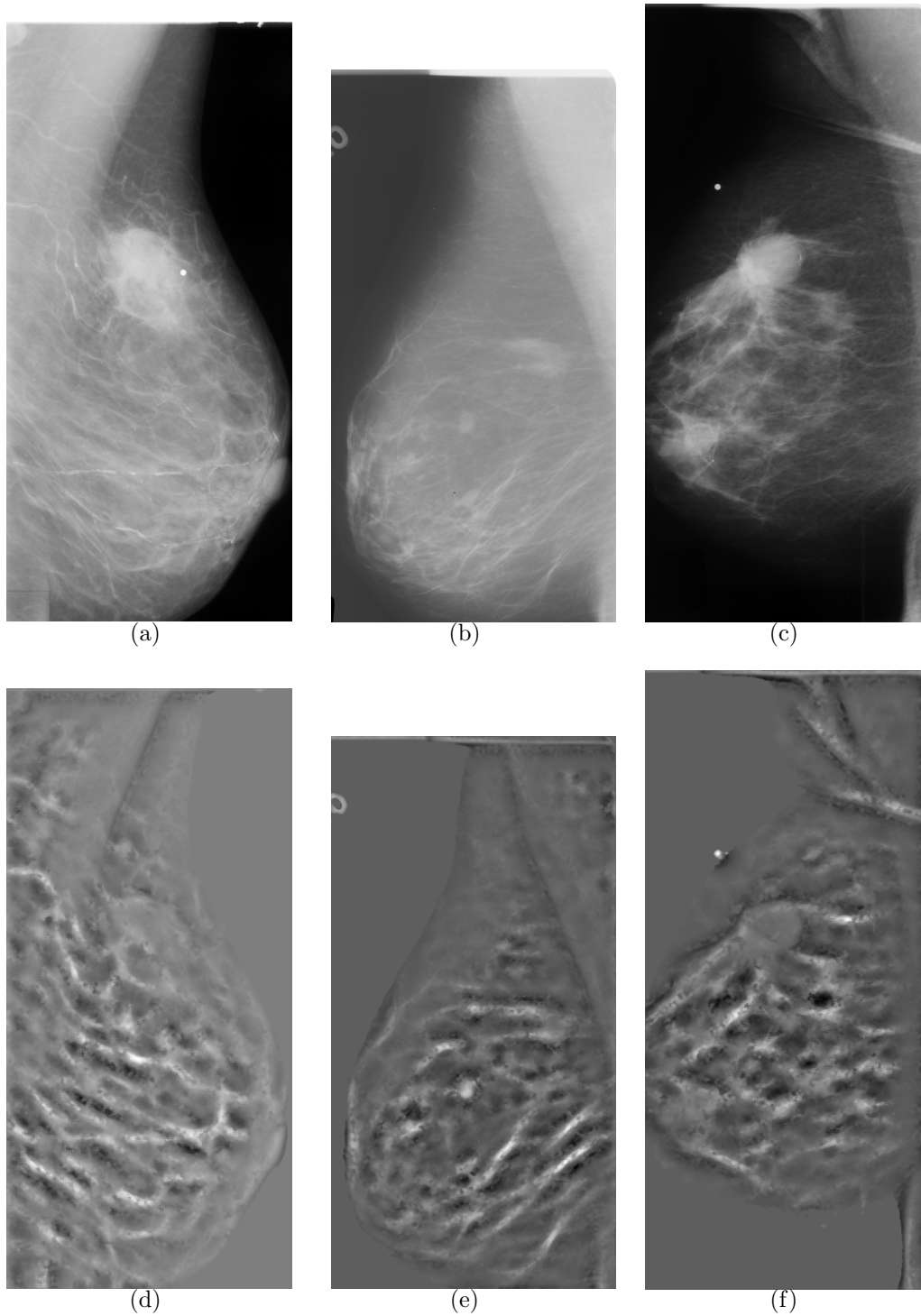


Figure 3.9: This figure shows examples of images in which the linear structures were enhanced using the oriented DoG filters. Figures 3.9(a), 3.9(b) and 3.9(c) show the original images and figures 3.9(d), 3.9(e) and 3.9(f) show the respective enhanced images.

CC and MLO view of each breast. Along with the digitized mammograms, the DDSM contains “boundary” files of the abnormalities. The outlines of the abnormalities are stored in “chain code” in these files. From this “chain code”, borders of the abnormalities can be reconstructed. After reconstructing the border, the interior of the outline is filled to create a “mask” of the mass. The mask can be regarded as the “ground truth”. In addition, images of spiculated lesions from the Mammography image analysis society (MIAS) database [79] were also used.

To report the performance of the detection algorithm, Free-response Receiver Operating Characteristic (FROC) curves ([18]) were generated for each set of images. A FROC curve is obtained by plotting sensitivity on the y-axis and the number of false positives per image on the x-axis. For this study, the different points on the FROC curve are obtained as follows. The N largest values in the magnitude response of all filter banks are marked as suspicious regions. The value of N was varied from 1 to 30. For each value of N , the sensitivity and the number of FPI were plotted to obtain the FROC curve.

3.5.2 Selection of Detection Algorithm Parameters Based on Measurement Data

Chapter 2 describes the measurement study conducted and the various parameters measured. These parameters include the length of the major axis of the central mass region, the width and length of all spicules, and the number

of spicules. In this section, we describe how the physical measurements are used to set the parameters of the detection algorithm. From our measurements of the physical parameters of spiculated masses and architectural distortions, we observed that the average width of spicules is 20 pixels. For computational efficiency, in this study we decimated the images by a factor of four and thus the average width of the spicules is approximately 5 pixels. Thus, to detect peaks, each column of the Radon domain is convolved with the following filter: $[-1 - 111111 - 1 - 1 - 1]^T$.

To account for variability in the shape and the structure of spiculated lesions, filter banks with different parameters are used. An estimate of the joint probability density function of the three variables (central mass region radius, length of spicules, and the number of spicules) was estimated. This was computed using a non-parametric density estimation method using Gaussian kernels. The joint estimate was then randomly sampled to select the parameters of the filter-banks. 25 filter-banks were created using these parameters and these are shown in Fig. 3.10.

In our preliminary studies, the algorithm marked the top N strongest matches per image. A pixel was only marked as detected if it was more than a fixed distance from the previously marked points; the distance was set as the average radius of the central region of a mass as measured in our measurement study. (This distance was 17 pixels for SM and 16 pixels for AD). By varying N , the tradeoffs in sensitivity and FPI can be visualized as an FROC curve.

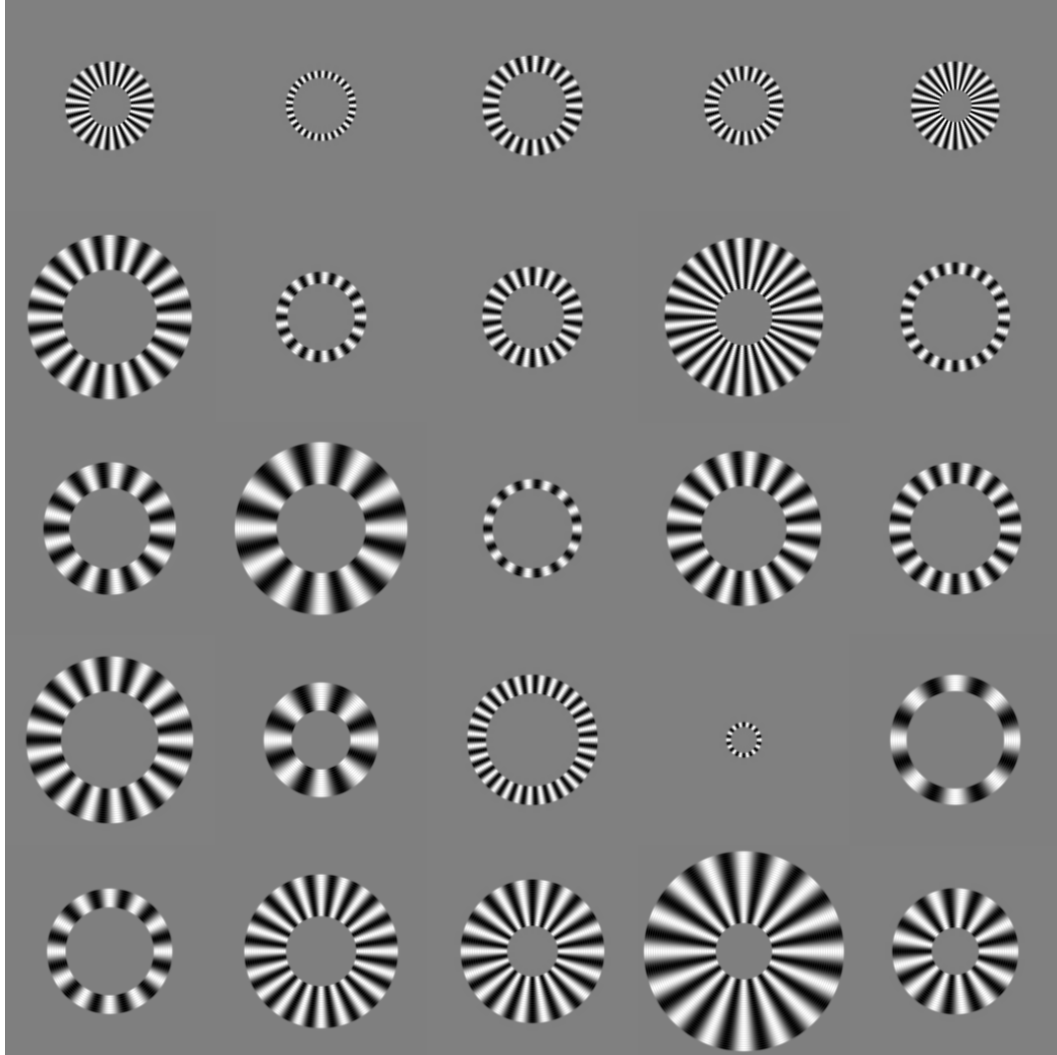


Figure 3.10: The set of 25 SFBs used for the detection of SM. To select the parameters of these SFBs an estimate of the joint pdf of the central mass region radius, length of spicules and number of spicules was obtained. This joint estimate of the pdf was then randomly sampled to select the parameters of the filter-banks.

3.6 Results

The detection algorithm described in the preceding sections were tested on various data-sets and the results were represented in the form of FROC curves. We tested this technique on multiple data-sets. The first two consisted of 50 spiculated masses and 46 architectural distortions from the DDSM database. The third set consisted of 19 spiculated masses from the Mammographic Image Analysis Society (MIAS) database ([79]) and finally the fourth set consisted of all malignant spiculated masses from the MIAS database. The third and fourth data-sets were specifically used as other researchers Liu *et al.* [22] and Karssemeijer *et al.* [21] have used these sets to test their algorithms on spiculated masses. Note that all of the measurement studies were conducted on a separate set of images from the DDSM database and there was no overlap between the measurement dataset and the testing dataset. The names of the all of the images used for the measurement study and for testing the performance of the detection algorithm are listed in Appendices A and B.

The results of applying the algorithm on the spiculated masses and architectural distortions from the DDSM dataset are shown in figure 3.11. For the spiculated masses, a sensitivity of 80% at 2.3 FPI was achieved and a sensitivity of 80% at 10.7 FPI was obtained for the architectural distortions.

The results on the set of images used by Karssemeijer *et al.* [21] are shown in figure 3.12(a). Karssemeijer achieved 90% sensitivity at 1 FPI and 100% sensitivity at more than 4 FPI. In comparison, we achieved a sensitivity of 90% at 1.3 FPI and a sensitivity of 100% at 2 FPI. Karssemeijer *et al.* also

included 31 normal images in the test set. If normal images are added to this set of spiculated masses, the number of false-positives per image will increase. The blue curve in figure 3.12(b) shows the FROC curve when the data-set contained images of both spiculated masses and 31 normal images. When these 31 normal images were included, a sensitivity of 90% at 2 FPI and a sensitivity of 100% at 2.8 FPI.

Similarly, figure 3.13(a) shows the results on the set of images used by Liu *et al.* [22]. Liu *et al.* achieved 84.2% sensitivity at 1 FPI and in comparison, we achieved a sensitivity of 84.2% at 3.1 FPI. Liu *et al.* also included 19 normal images in the test set. The blue curve in figure 3.13(b) shows the FROC curve when the data-set contained images of both spiculated masses and 19 normal images. When these 19 normal images were included, a sensitivity of 84.2% at 3.5 FPI was achieved.

3.7 Discussion

3.7.1 Comparison with previous studies

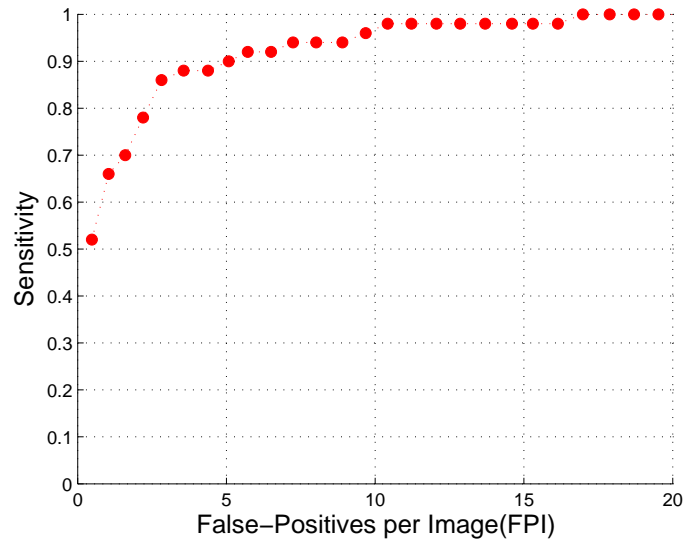
In this section we compare the performance of our algorithm with previous studies. Table 3.1 summarizes the results of prior studies for the detection of the spiculated masses and/or architectural distortions. Table 3.1 shows the number of training and testing images used in each study and the number of spiculated masses and/or architectural distortions present in the training and testing sets. From this table, we see that the number of spiculated masses used to test the detection algorithm is quite small. The major disadvantage

Author	Total # of images	Total # of SM & AD	# Training Images (# of SM/AD)	# Testing Images (# of SM/AD)	Sens	FPI
Kegelmeyer [20]	330	68	165(34)	165(34)	100	–
Karssemeijer [21]	50	19	none	50(19)	90	1
Delp [22]	38	19	19(9)	19(10)	84.2	1
Zwiggelaar [80]	54	27	Leave-one-out classification method used		70	0.01

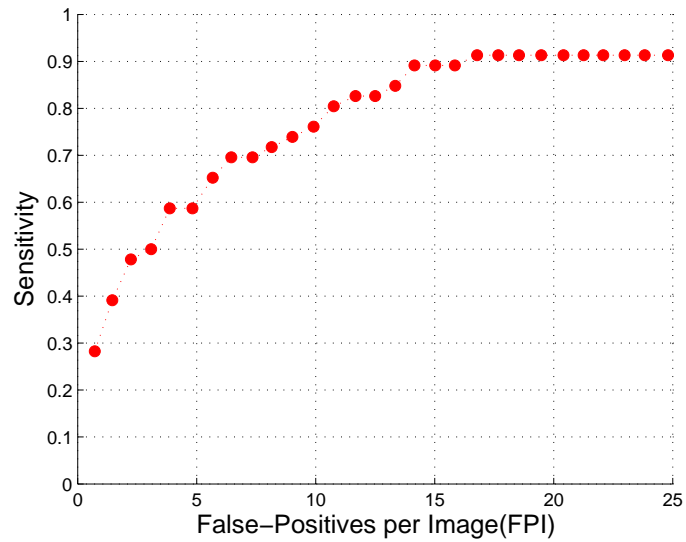
Table 3.1: This table summarizes the performance of prior algorithms developed for the detection of spiculated masses and architectural distortions. It also presents the number of training and testing images used in each study.

of testing an algorithm on a small set of images is that the natural variations among the spiculated masses and architectural distortions are not seen by the algorithm.

To the best of our knowledge the experiments carried out in this dissertation are the largest detection studies (in terms of the number of images of spiculated lesions used) on any database. The accuracy of our detection algorithm may be improved further by including a classifier to differentiate between the spiculated lesions and normal tissue.

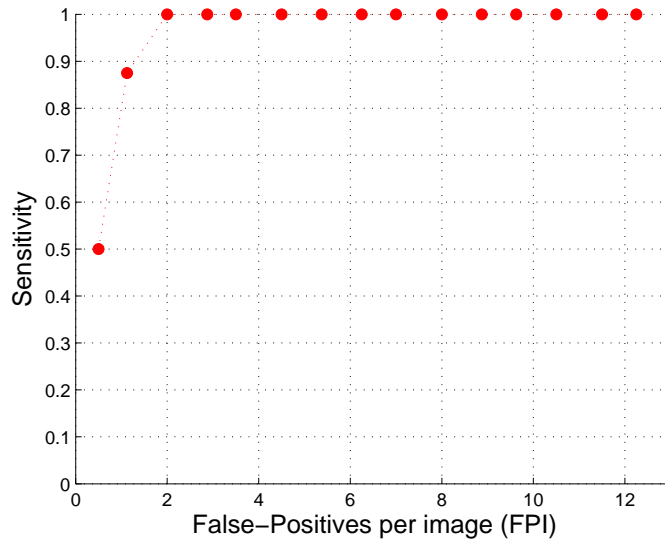


(a) FROC curve: Spiculated Masses

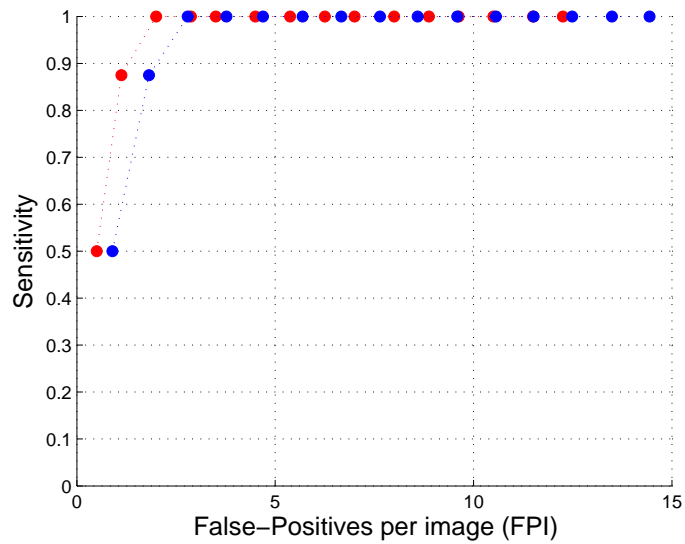


(b) FROC curve: Architectural Distortions

Figure 3.11: FROC Curves: The performance of the detection algorithm is reported using FROC curves. The y-axis of the FROC curve corresponds to the sensitivity of the algorithm and the x-axis corresponds to the number of FPI per image. Figures 3.11(a) and 3.11(b) shows the FROC curve for the set of 50 spiculated masses and 46 architectural distortions from the DDSM database, respectively.

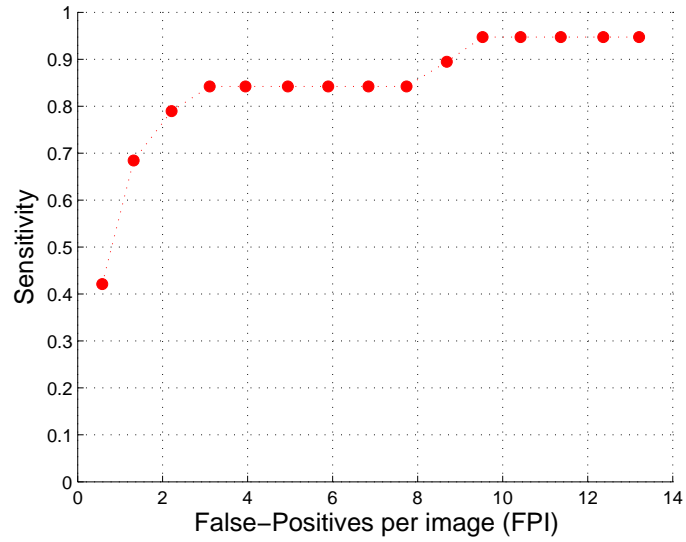


(a) FROC curve for spiculated masses only

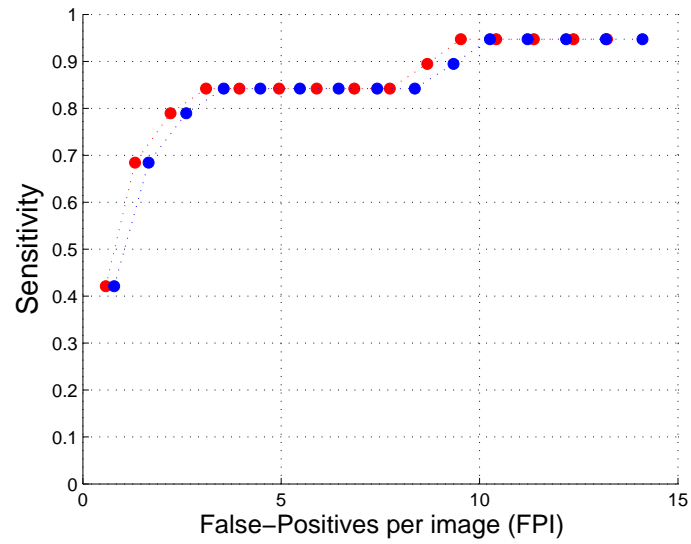


(b) FROC curves for spiculated masses with normal images

Figure 3.12: FROC Curves: Figure 3.12(a) shows the results on the set of images used by Karssemeijer *et al.*. If normal images are added to this set of spiculated masses, the number of false-positives per image will increase. The blue curve in figure 3.12(b) shows the FROC curve when the data-set contained images of both spiculated masses and 31 normal images.



(a) FROC curve for spiculated masses only



(b) FROC curves for spiculated masses with normal images

Figure 3.13: FROC Curves: Figure 3.13(a) shows the results on the set of images used by Liu *et al.*. If normal images are added to this set of spiculated masses, the number of false-positives per image will increase. The blue curve in figure 3.13(b) shows the FROC curve when the data-set contained images of both spiculated masses and 19 normal images.

Chapter 4

Comparison of Spicule Enhancement Algorithms

4.1 Introduction

In Chapter 3, (section 3.2.1) a new algorithm for the enhancement of spicules was described. In this method, the Discrete Radon Transform (DRT) of an image is computed and a filter is applied in the Radon domain. The enhanced image is obtained by computing the Inverse DRT.

Recently, the Fast Slant Stack (FSS) method ([81]) was developed to compute the DRT of an image. This algorithm is one-to-one and is invertible on its range, and it is computationally efficient. In this chapter, we compare the effect of using the traditional implementation of the DRT versus the FSS implementation of the DRT in our spicule enhancement algorithm.

4.2 Materials and Methods

4.2.1 Brief review of spicule enhancement algorithm

In this section, we provide a description and the intuition behind the Radon-based approach to enhancing spicules. The most natural mathematical framework for enhancing and detecting straight line structures in images is the

Radon Transform [69]. The Discrete Radon Transform $\hat{g}(\rho, \theta)$, of an N-by-N image $f(x, y)$ is given by:

$$\hat{g}(\rho, \theta) = \sum_{y=-N/2}^{N/2} \sum_{x=-N/2}^{N/2} f(x, y) \delta(\rho - x \cos(\theta) - y \sin(\theta)) \quad (4.1)$$

Where $\delta(r)$ is the Kronecker impulse function, which is zero everywhere except when $r = 0$. Hence, the term $\delta[\rho - x \cos(\theta) - y \sin(\theta)]$ contributes summed values of $f(x, y)$ only along the line $[\rho - x \cos(\theta) - y \sin(\theta)]$ and thus the value of $\hat{g}(\rho, \theta)$ for any (ρ, θ) is the sum of values of $f(x, y)$ along this line. An important property of the Radon Transform is that a line in the image space $f(x, y)$ maps to a unique peak in the Radon domain. In addition, lines of different thickness have different representations in the Radon domain.

Figure 3.2 in Chapter 3, shows two lines of different thickness and their corresponding Radon Transforms. A single pixel thick line would be represented by a point in the Radon domain whereas a 4 pixel thick line would be represented by 4 points along a column in the Radon domain. Thus, by detecting “local peaks” along the columns in the Radon domain, it is possible to detect the corresponding lines in the image. To do so, a number of peak detection algorithms can be used. While many approaches might be taken to identify peaks, our approach is to simply filter the rows of the Radon Transform with an appropriate bi-valued one-dimensional linear filter with an impulse response that is a rectangle of width agreeing with the maximum spicule width to be highlighted. This approach has the advantages of both directness and simplicity. Application of this simple filter to the Radon image, followed by

an Inverse Radon Transform, will yield an enhanced image with amplified linear structures (spicules) of the requisite widths, all other structures being suppressed.

4.2.2 Experimental Setup

The images for this study were obtained from the Digital Database for Screening Mammography (DDSM), ([62]). Thirty MLO images, each containing a single spiculated mass were selected. A region-of-interest containing the abnormality was cropped and used for all further analysis. The 30 images were enhanced using two versions of our spicule enhancement algorithm. In the first version (Method 1), we used the traditional implementation of the DRT and in the second version (Method 2), we used the FSS implementation of the DRT. To study the advantages of using the FSS over the traditional implementation of the DRT, we followed the following procedure. We conducted multiple 2-alternative-forced choice (2 AFC) observer studies. In the first study, the observer was an expert radiologist (R1) with extensive expertise in breast imaging. In the second observer study, 10 engineering students from our research lab were selected as observers. For each image, the original image and the two enhanced images (from Method 1 and Method 2) were shown to all of the observers and they were asked to choose the image in which the spicules were enhanced more prominently. We also measured the total time for obtaining 30 enhanced images using methods 1 and 2. Both methods were implemented in MATLAB (The MathWorks, Natick, MA). Figure 4.1 shows

this experimental setup. The image in the center is the original image. In this example, the enhanced image obtained with Method 1 is shown on the left and the enhanced image obtained with Method 2 is shown on the right. The positions of the enhanced images (left or right) were randomly selected so that the observers would not be able to tell by which method enhanced image was generated.

4.2.3 Results

The results of the 2 alternative-forced choice (2 AFC) observer study are presented in Table 4.1. For each observer, we counted the number of images for which he/she felt that the spicules were more apparent in the image enhanced by Method 2 as compared to the image enhanced by Method 1. The radiologist preferred the enhanced images generated by Method 2 for 28 out of the 30 images (93.33%). Moreover, most of the engineering observers also preferred the enhanced images of Method 2 over that of Method 1. On average, for 74.6% (standard deviation = 21.1%) of the images all of 10 engineering observers felt that enhancement results of Method 2 were visually more appealing. There were only two images for which the radiologist, as well as most of the novice observers, preferred the enhanced image generated by Method 1 to that generated by Method 2 (Fig. 4.2). This was most likely due to the fact that the radiologist reported that he preferred images where edges were enhanced. He felt he could see more spicules in Figure 4.2(c) as compared to Figure 4.2(a). In addition, he could see more detail in the region of the mass

Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5	Obs. 6	Obs. 7	Obs. 8	Obs. 9	Obs. 10
93% (28/ 30)	47% (14/ 30)	87% (26/ 30)	67% (20/ 30)	93% (28/ 30)	70% (21/ 30)	100% (30/ 30)	60% (18/ 30)	90% (27/ 30)	40% (12/ 30)

Table 4.1: This table shows the results of the two-forced alternative choice (2-AFC) experiment. For each observer, the percentage of images where the observer felt the spicules were enhanced more prominently by Method 2 (which used the FSS technique) is shown. On average, for 74% of images the observers felt that Method 2 produced better enhancement results. In comparison, the radiologist preferred the images generated by Method 2, for 28 out of the 30 images (98%).

in Figure 4.2(d) as compared to Figure 4.2(f). Similarly, three examples of images where most observers preferred the enhanced image generated by Method 2 are shown in Figure 4.3. Method 2 was computationally more efficient than Method 1. The total time required to run Method 1 on 30 images was 29.5 minutes whereas the total time required running Method 2 on 30 images was only 4.7 minutes. Thus, Method 2 was six times faster than Method 1.

4.3 Discussion

In Chapter 3, a novel algorithm for the CADe of spiculated lesions is described and an important component of this algorithm is a spicule enhancement strategy. In this method, we employ the DRT of an image and filter in the Radon domain. In this chapter, we have compared the effect of using the traditional implementation of the DRT (Method 1) versus the FSS implementation of the DRT (Method 2) in our spicule enhancement algorithm. The

results show that observers with and without experience in radiology found images enhanced by Method 2 to be more visually appealing. Moreover, Method 2 was computationally more efficient than Method 1 as it utilized the FSS implementation of the DRT whereas, Method 1 utilized the traditional implementation of the DRT algorithm. In future work, additional observer studies could be conducted to analyze the effect of the enhancement algorithms on detection performance. One could assess if the enhancement techniques help observers to detect more spiculated lesions. The effect of the enhancement methods on CADe algorithms could also be studied. The initial evaluation of CADe algorithms is typically in terms of FROC curves [82] of the algorithms acting independently (i.e., no human observer). However, as CADe algorithms are ultimately used to assist radiologists, confirmatory studies of the performance of the radiologist with and without the use of CADe would also be needed.

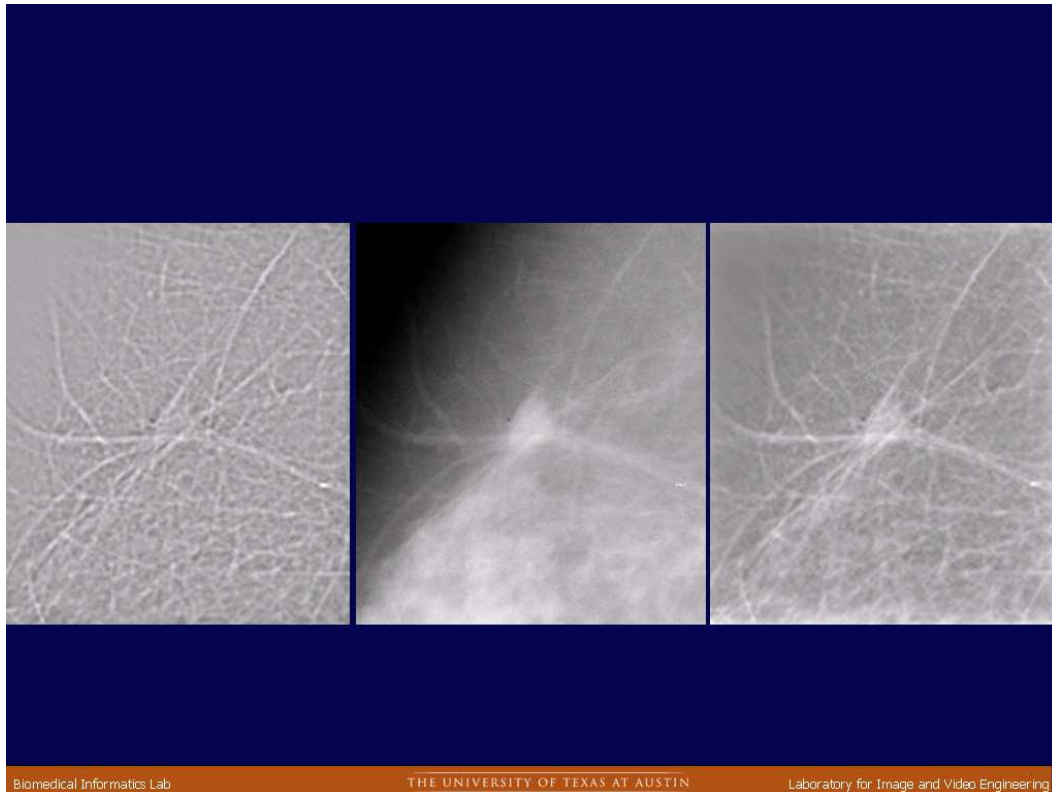


Figure 4.1: This figure shows a screenshot of the observer experiments carried out. The image in the center is the original image and the images on the left and right are the enhanced images obtained with the two enhancement methods. The observer is asked to choose which enhanced image they found most visually appealing. For the 2-AFC observer studies experiments the observers were not told from which method a particular enhanced image was generated and the images were shown in random order.

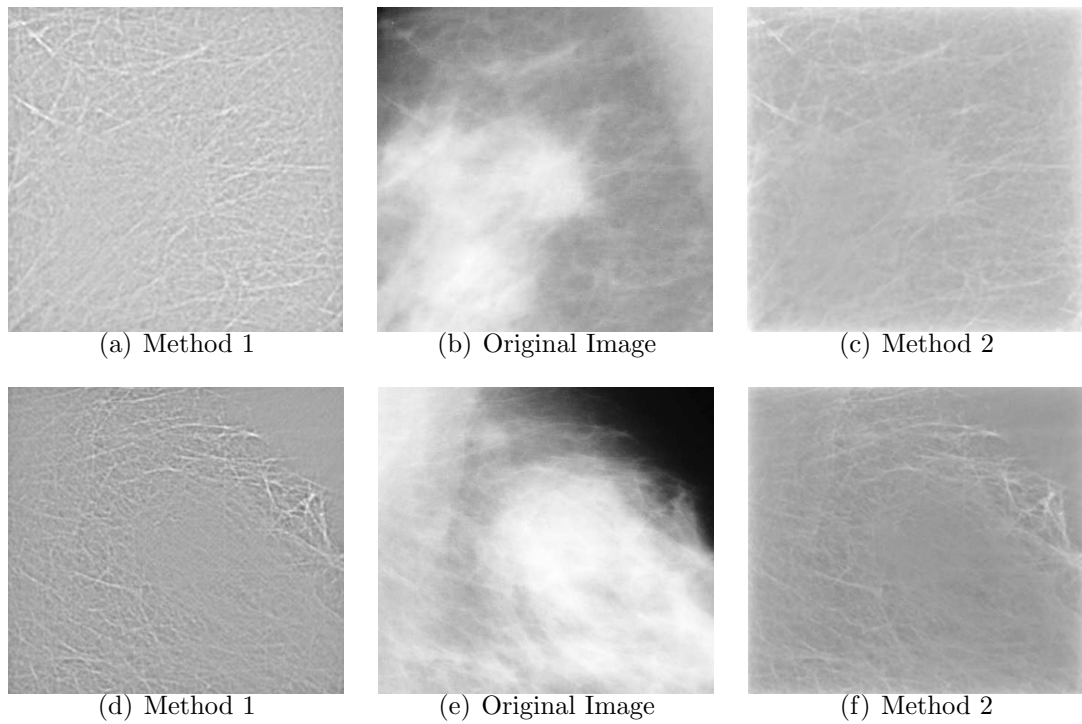


Figure 4.2: The only two images where most of the readers liked Method 1 better than Method 2. These were also the only two images where the radiologist found the enhanced image created with Method 1 more visually appealing.

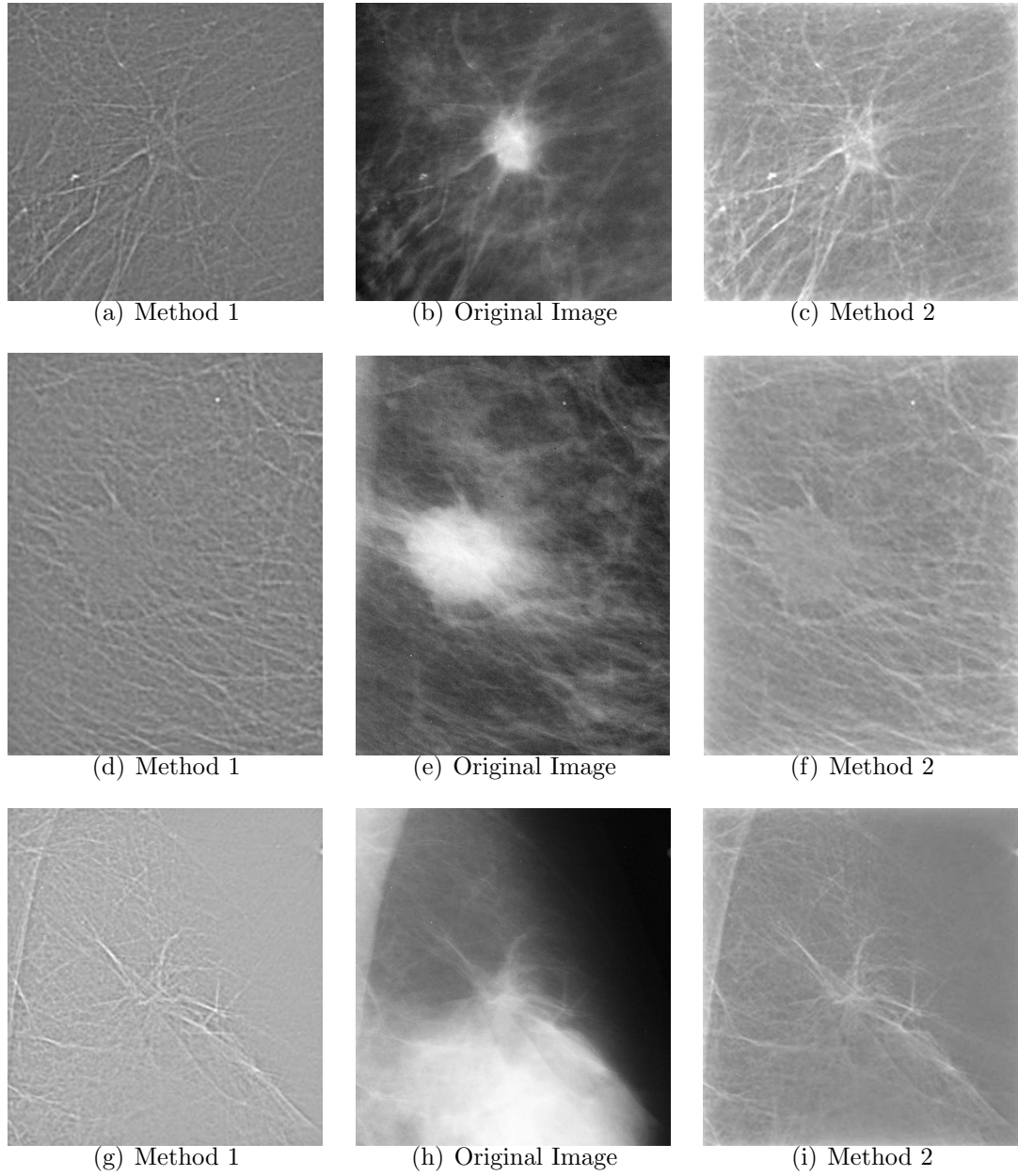


Figure 4.3: Three images where most of the readers liked the images enhanced with Method 2 better than the images enhanced with Method 1.

Chapter 5

Similarity Metrics for Binary Images: Comparing Segmentations on Medical Imaging

5.1 Introduction

Accurate image segmentation is critical for the detection and classification of objects-of-interest in medical images. Image segmentations may be generated by human observers or by computer algorithms. In Chapter 2, it was shown how various statistical techniques such as ICC were used to quantify inter- and intra-observer variability in the measurement of the physical characteristics of spiculated lesions. During the course of this work it was observed that the existing statistical techniques have a limitation when used to compare segmentations of linear structures. Specifically, they do not account for the spatial location of these linear structures.

In this chapter, similarity metrics were empirically compared, with an emphasis on their value for comparing segmentations of linear structures (e.g., spicules on mammographic lesions). Intensity-based similarity metrics (e.g., Dice) assume that the two images being compared are registered and so their similarity is determined from a comparison of the intensity associated with pixels with the same coordinates in the two images. Spatial-based methods

(e.g., PCM) first match pixels between the two images based on intensity and the similarity of the two images is determined from a comparison of the spatial coordinates of the corresponding pixels in the two images. Alternatively, one could design metrics in the wavelet domain that are less sensitive to small geometric distortions such as translation and rotation. The Complex-Wavelet Structural Similarity Index (CW-SSIM) was originally developed for gray-scale image quality assessment and it has not been previously used in medical imaging. The key idea behind the CW-SSIM method is that certain image distortions lead to consistent phase changes in the local wavelet coefficients and that a consistent phase shift of the coefficients does not change the structural content of the image.

A segmentation is represented as a binary image where all pixels on and inside the indicated boundary are assigned a value of one and zero otherwise. For dense objects, e.g., tumors, the segmentation may contain many pixels. By comparison, for linear structures, e.g., blood-vessels, the segmentation can be sparse. Similarly, small dense objects are represented by a few pixels (sparse segmentations). For example, consider a spiculated mass as seen on mammography (Fig. 5.1(a)). In Fig. 5.1(b) the central region of the mass, a dense object, has been segmented, whereas in Fig. 5.1(c) the spicules of the mass, linear structures, have been segmented.

Segmentation generators are evaluated in terms of three qualities: precision (reproducibility), accuracy (agreement with truth), and efficiency (time taken) [83]. A surrogate of truth and a metric for assessing the similarity be-

tween it and a given segmentation are needed to assess accuracy. Similarity metrics are also needed to quantify the variability between image segmentations created by different segmentation generators. An image is a function $f(x, y)$ of intensity over spatial coordinates (x, y) . Thus, a similarity metric for images in this representation must account for similarity in both intensity and space, and metrics can be classified according to how they handle these two properties. One class of approaches assumes that the two images being compared are registered and so their similarity is determined from a comparison of the intensity associated with pixels with the same (x, y) coordinates in the two images. We will refer to such metrics as “intensity-based”. Alternatively, pixels may first be matched between the two images based on intensity and the similarity of the two images is determined from a comparison of the spatial coordinates of the corresponding pixels in the two images. We will refer to such metrics as “spatial-based”.

When evaluating the similarity of segmentations of dense objects, it is natural to think of the degree of the spatial overlap, and so metrics for the evaluation of segmentations of dense objects tend to be “intensity-based”. In comparison, the similarity of two segmentations of linear structures can be expressed in terms of the spatial distances between linear structures in the two images. Thus, metrics for the evaluation of segmentations of linear structures tend to be “spatial-based”. In medical imaging the segmentation of dense objects is encountered more frequently and so many metrics have been developed for the comparison of segmentations of dense objects. However, few

metrics have been proposed for or tested on segmentations of linear structures in medical images.

We will now summarize the intensity-based metrics that are commonly used in medical image processing. Since we are concerned with image segmentations that are represented as binary images, the intensity $f(x, y)$ can only take on the values 0 or 1. Thus, the simplest of the intensity-based metrics quantify “spatial overlap” by applying Boolean operations to the corresponding pixel intensities. Examples of these straightforward and easy to implement metrics include the Dice [84], Jaccard [85], Simpson [86], Ochiai [87], Braun-Blanquet [88], and Sokal-Sneath [89] similarity metrics. Among them, the Dice similarity coefficient [84] and the Jaccard metric [85] are widely used in medical imaging. Notice that as a consequence of their design, these metrics penalize segmentations that are different by even just one pixel. While this property is probably desirable for applications where the goal is to compare segmentations of dense objects encompassing many pixels, it is of questionable value for assessing the similarity of tracings of linear structures comprised of relatively few pixels.

Another commonly used intensity-based similarity metric is the Mean Square Error (MSE) [69]. It is computed by averaging the squared intensity differences of the two images. Note that while other intensity-based metrics report agreement between two images, the MSE is designed as a measure of the error or disagreement. In general, the MSE is zero when the two images are identical and has no upper bound. By comparison, most intensity-based

metrics are designed so that they have an upper-bound of one and a lower-bound of zero (representing completely dissimilar images). For binary images, it can be shown that the MSE is a linear function of the Simple Matching metric [90] (Table 5.3): $MSE = 1 - Simple\ Matching$.

The Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm was recently developed by Warfield *et al.* [91] for the evaluation of dense object segmentations. Given a set of segmentations (at least 3-5) of the same object, the STAPLE algorithm provides a robust statistical framework which simultaneously and iteratively generates statistical estimates of the ground truth and the accuracy of each segmentation in terms of sensitivity and specificity. The STAPLE algorithm is unique in that it provides a probabilistic estimate of the hidden ground truth. The accuracy of each segmentation is measured in terms of sensitivity and specificity. The sensitivity is defined as the fraction of the pixels labeled as one in the ground truth that are also labeled as one in the segmentation being assessed. The specificity is the fraction of the pixels labeled as zero in the ground truth that are also labeled as zero in the segmentation being assessed [91]. Thus, the STAPLE algorithm uses intensity-based similarity metrics to compare each segmentation to the ground truth.

All intensity-based metrics discussed so far use a pixel-level comparison to compute a measure of similarity. They do not take advantage of the “local” information around each pixel. The recently developed Structural Similarity (SSIM) index proposed by Zhou *et al.* [92] uses the “local” structural informa-

tion to define a similarity metric. The structural information in an image is defined as those attributes that represent the structure of objects in the scene, independent of the mean intensity and contrast. Thus, the SSIM metric compares local patterns of pixel intensities that have been normalized for mean intensity and contrast. It has been shown that the SSIM metric is superior to existing metrics used for image quality assessment [92]. To the best of our knowledge, SSIM has not been used in medical imaging applications.

Intensity-based metrics have limitations when observers outline/trace linear structures. It is very likely that outlines which represent the same structures, but were traced by different observers, are a few pixels apart. However, intensity-based metrics do not account for “spatial proximity” in these situations. One possible approach to overcome this limitation is to dilate the the outline images before computing an intensity-based metric. For example, Jomier *et al.* proposed a method in which they combine the STAPLE algorithm with morphological operators (dilation) to compare segmentations of vascular structures (e.g., vessels) [93].

Alternatively, one could compare segmentations of linear structures using spatial-based similarity metrics that have been proposed in the image processing literature to quantify the similarity of edge images. “Edges” are characterized by sharp transitions in intensity between two regions and segmentations of linear structures can be viewed as edges of a predefined width. Since these metrics assess spatial proximity rather than assuming images are registered in space, they are better suited than intensity-based metrics for

comparing segmentations of linear structures. To the best of our knowledge, these metrics have not previously been used for medical imaging applications.

Spatial-based similarity metrics include the Figure of Merit (FOM) [94], Closest Distance Metric (CDM) [95, 96], and Pixel Correspondence Metric (PCM) [96]. All of these metrics allow for a small localization error (user-specified) and so an edge pixel does not have to be at the exact same spatial location as the ground truth pixel to be classified as a correct segmentation. In essence, each of the pixels in the edge image has to be “matched” or “mapped” with a pixel in the ground truth image. Some metrics (e.g., FOM) allow more than one detected edge pixel to correspond to the same ground truth pixel (many-to-one mapping). This is undesirable and thus CDM and PCM were designed such that a one-to-one mapping between pixels in the edge image and the ground truth is enforced. Prieto *et al.* reported that PCM produces a more accurate matching than CDM [96]. Finally, while most spatial-based metrics account only for the distance of edges, a few metrics (e.g., PCM) also account for the “strength” or intensity of edges (this has an advantage for applications in which gray-scale images are compared).

All metrics discussed above are designed in the spatial domain. Some assume that the images are registered and compute a metric based on image intensities. Otherwise, if the images are not registered, then a matching step is required where each pixel in one image is matched to a pixel in the ground truth image. This process can be computationally intensive.

Alternatively, one could design metrics that are less sensitive to geo-

metric distortions. It is often easier to perform such analyses in the Fourier or Wavelet domains. For example, translation and rotation invariant shape descriptors have been defined in the Fourier domain [97]. Developing a metric in the Wavelet domain (rather than the Fourier domain) has numerous advantages, since the Wavelet domain provides localization in both space and frequency.

SSIM has been extended to the complex-wavelet domain [98]. The key idea behind the CW-SSIM method is that certain image distortions lead to consistent phase changes in the local wavelet coefficients and that a consistent phase shift of the coefficients does not change the structural content of the image. The CW-SSIM index analyzes phase consistency to achieve robustness to small geometric distortions. CW-SSIM was originally developed for general-purpose gray-scale image quality assessment and pattern recognition applications and it has not been previously used in medical imaging applications. We show that CW-SSIM is a good metric for assessing similarity of segmentations of linear structures because it is robust to small geometric distortions. Whether a metric is robust to small geometric distortions is probably not important when segmentations of large, dense objects are compared, but may become crucial when segmentations of linear structures are assessed. For example, consider the case when an expert twice traces a linear structure in an image and the resulting two outlines are displaced by a single pixel. The expert may have intended these to be outlines of the same structure and visually the two measurements may appear very similar. However, an intensity-based

similarity metric would indicate that the two segmentations are very different. In principle, spatial-based metrics are better suited to quantify the similarity of measurements of linear structures. However, spatial-based metrics tend to have high computational complexity and require parameters to be set upfront.

The purpose of this study was to investigate metrics for assessing the similarity of segmentations, with an emphasis on their value for comparing segmentations of linear structures. For this study, the metrics have been tested on two sets of data. The first set consists of simulated images to which geometric transformations like rotation and translation were applied. The second set consists of outlines of linear structures on images from mammography.

5.2 Metrics

5.2.1 Intensity-Based Similarity Metrics

The simplest of the intensity-based similarity metrics are based on the direct measurement of spatial overlap of segmentations. Consider the case when two segmentations of an object, created by different segmentation generators, are to be compared. Let a represent the number of pixels that are non-zero in both binary segmentation images. Note that, the greater the spatial overlap, the greater the value of the term a . The term c represents the number of pixels marked as one by segmentation generator #1 only and b represents the number of pixels marked as one by segmentation generator #2 only. Finally, d represents the number of pixels that were labeled as zero by both segmentation generators. The terms a , b , c , and d form a two-by-two con-

tingency table as shown in Table 5.1. Metrics designed to measure the spatial overlap between two segmentations are computed by dividing the term a by a normalizing factor so that the values are in the range 0 to 1. The normalizing factor is formed by a combination of the terms a , b , c and d . A compre-

Segmentation 2	Segmentation 1			
		One	Zero	Total
	One	a	b	a+b
	Zero	c	d	c+d
	Total	a+c	b+d	a+b+c+d

Table 5.1: Two-by-two contingency table: The goal is to compare two segmentations of the same object. A pixel that lies on or inside the boundary of the segmented object is labeled one and zero otherwise. In this table a represents the number of pixels that are non-zero in both images. The term b represents the number of pixels marked as one by segmentation generator #1 only and c represents the number of pixels marked as one by segmentation generator #2 only. Finally, d represents the number of pixels that were labeled as zero by both segmentation generators.

hensive review of spatial-overlap metrics was published by Shi [99]. Several spatial-overlap based metrics are summarized in tables 5.2 and 5.3. The Dice and Jaccard metrics are commonly used in medical imaging, while the other metrics (Tables 5.2 and 5.3) are routinely used in other fields [99–101].

From a conceptual viewpoint, spatial overlap metrics can be divided into two categories. The metrics in Table 5.2 do not use the term d (Table 5.1) whereas the metrics in Table 5.3 do. For purposes of evaluating similarity of segmentations, the measures in Table 5.2 are more intuitively appealing than those in Table 5.3 because d represents the number of pixels that were labeled as zero in both segmentations. In other words, most users would probably

prefer that a similarity measure for image segmentations be independent of the size of the background. Of course, the metrics in Table 5.3 may be suitable for other applications where both agreement variables (a and d) are equally informative.

The Dice similarity coefficient [84] is commonly used in medical imaging studies to quantify the degree of overlap between two segmented objects, e.g., [102–104]. In terms of the quantities in Table 5.1, Dice is defined as:

$$\text{Dice} = \frac{(2 \times a)}{(2 \times a + b + c)} \quad (5.1)$$

If the two segmentations overlap completely then $\text{Dice} = 1$ and if there is no overlap then $\text{Dice} = 0$. Although a Dice value of greater than 0.7 is taken to indicate “excellent” agreement [103] it is difficult to analyze the absolute value of Dice since “similarity” is application dependent and represents different things in different contexts.

Another metric commonly used in medical imaging is the Jaccard metric [85]. However, the Jaccard and Dice metrics are closely related: $\text{Dice} = (2 \times \text{Jaccard})/(\text{Jaccard} + 1)$. Thus, the Jaccard metric does not provide independent information in addition to the Dice metric and so we do not compute both metrics in this study. Moreover, since all the metrics listed in Table 5.2 are all closely related, some are equivalent under certain circumstances. For example, if ($b = c$) then the Dice, Kulczynski2, Braun-Blanquet, Simpson, and Ochiai metrics are equivalent and simplify to the formula $a/(a + b)$. This

example corresponds to the case in which the total number of pixels marked in each of the two segmentations are equal.

Name of Metric	Source and Year	Formula	Range
Dice	Dice, 1945 [84]	$\frac{2 \cdot a}{2 \cdot a + b + c}$	0 to 1
Jaccard	Jaccard, 1912 [85]	$\frac{a}{a + b + c}$	0 to 1
Kulczynski1	Kulczynski, 1928 [105]	$\frac{a}{b + c}$	0 to inf
Kulczynski2	Kulczynski, 1928 [105]	$\frac{(a/2) \cdot (2 \cdot a + b + c)}{(a + b) \cdot (a + c)}$	0 to 1
Simpson	Simpson, 1960 [86]	$\frac{a}{\min(a + b, a + c)}$	0 to 1
Ochiai	Ochiai, 1957 [87]	$\frac{a}{((a + b) \cdot (a + c))^{1/2}}$	0 to 1
McConnaughey	Hubalek, 1982 [106]	$\frac{a^2 - b \cdot c}{(a + b) \cdot (a + c)}$	-1 to 1
Braun-Blanquet	Braun and Blanquet, 1932[88]	$\frac{a}{\max(a + b, a + c)}$	0 to 1
Sokal & Sneath(2)	Sokal & Sneath 1963 [89]	$\frac{a}{a + 2 \cdot b + 2 \cdot c}$	0 to 1

Table 5.2: Similarity metrics for binary data: These are a set of metrics that are defined based on the terms a , b , and c only (from Table 5.1). The Dice and Jaccard metrics have been commonly used in medical imaging applications whereas the remaining metrics are more popular in other fields. These metrics are all closely related.

5.2.2 Spatial-Based Similarity Metrics

In applications where readers outline/trace linear structures, the binary representations of the segmentation can be modeled as edges in the image. Thus, another group of relevant metrics are those that were originally

Name of Metric	Source and Year	Formula	Range
Russell and Rao	Russel and Rao, 1940 [107]	$\frac{a}{a + b + c + d}$	0 to 1
Simple Matching	Sokal & Michener, 1958 [90]	$\frac{a + d}{a + b + c + d}$	0 to 1
Yule	Yule, 1900 [108]	$\frac{(a \cdot d) - (b \cdot c)}{(a \cdot d) + (b \cdot c)}$	0 to 1
Rogers and Tanimoto	Rogers & Tanimoto, 1960 [109]	$\frac{a + d}{a + d + 2 \cdot (b + c)}$	0 to 1
Sokal and Sneath (1)	Sokal & Sneath, 1963 [89]	$\frac{2 \cdot (a + d)}{2 \cdot (a + d) + b + c}$	0 to 1

Table 5.3: Metrics for binary data: These are a set of metrics that are defined based on the terms a , b , c , and d (from Table 5.1). Note that for image segmentation similarity, d represents the number of pixels that were labeled as zero in both segmentations. Thus, these metrics may not be as desirable as those in Table 5.1 since they depend on the size of the background surrounding the segmented objects.

proposed in the image processing literature to quantify the similarity of edge images. These include the Figure of Merit (FOM,[94]), Closest Distance Metric (CDM,[95]), and Pixel Correspondence Metric (PCM,[96]). Edge-based similarity metrics are suitable for assessing the similarity of outlines of linear structures on medical images since they take into account the spatial proximity of the outlines created. In this study, FOM was not used because it has a number of limitations [96]. The major limitation of the FOM metric is that it allows more than one edge pixel in one image to match or correspond to the same pixel in the other image and thus may provide an unrealistic measure of image similarity. PCM and CDM avoid this issue by forcing a one-to-one matching between the edge pixels in the two images. The difference between

PCM and CDM lies in the manner in which the matching is done. While CDM uses a simple closest distance measure, PCM uses a sophisticated technique from graph theory.

5.2.2.1 Pixel Correspondence Metric (PCM)

The PCM was proposed by Prieto *et al.* [96]. Let f and g represent two images that are to be compared. Let $f(i, j)$ represent a particular pixel in f and $g(k, l)$ represent the corresponding pixel in g . For every pixel $f(i, j)$, the PCM algorithm tries to find its match within a neighborhood of radius η of the corresponding pixel $g(k, l)$. This idea is shown schematically in Fig. 5.2. The metric is given by the following equation:

$$PCM_{\eta}(f, g) = 100 \cdot \left(1 - \frac{C(M_{opt}(f, g))}{|f \cup g|} \right) \quad (5.2)$$

where $C(M_{opt}(f, g))$ is the cost of optimal matching between the images, η denotes the localization error allowed between the pixels and $|f \cup g|$ is the total number of pixels that are not zero in f or g . (Note that in in terms of the quantities described in Table 5.1, $|f \cup g| = a + b + c$). The search for the optimal matching is a complex problem and is solved by modeling the task as a weighted matching task in bipartite graphs. A number of solutions have been proposed. A detailed description of the PCM algorithm can be obtained in [96].

5.2.2.2 Closest Distance Metric (CDM)

The CDM was proposed by Bowyer *et al.* [95]. Let f and g denote the two images. The goal is to find a possible match in g (within a certain neighborhood), for every pixel $f(i, j)$ in images f . If multiple matches exist, then the edge pixel $g(k, l)$ that is closest to $f(i, j)$ is selected. Finally, the number of matched and unmatched pixels are counted. The metric is computed as:

$$CDM_{\eta}(f, g) = 100 \cdot \left(1 - \frac{C(M_{cd}(f, g))}{|f \cup g|} \right) \quad (5.3)$$

where $CDM_{\eta}(f, g)$ is the cost of the matching obtained using the closest-distance condition, η is the neighborhood radius used in the matching, and $|f \cup g|$ is the total number of pixels that are not zero in f or g . (Note that in terms of the quantities described in Table 5.1, $|f \cup g| = a + b + c$).

5.2.3 Complex Wavelet Structural Similarity (CW-SSIM)

The structural similarity (SSIM) index was originally proposed for the prediction of human preferences in evaluating image quality [92]. The underlying assumption of this approach is that the human visual system (HVS) is highly adapted to extract structural information from the visual scene and thus a measure of structural similarity should provide a good estimate of perceived image quality. It has been demonstrated that the SSIM index is successful in predicting the quality of images degraded with a wide variety of distortion types and levels. However, the complex wavelet SSIM (CW-SSIM) index [98] has proven to be more robust than the baseline SSIM index to small geometric

distortions. In previous work, both the baseline SSIM index and the CW-SSIM index were developed and tested for the comparison of gray-scale images only. Here we propose to use the CW-SSIM index as a binary image similarity metric, where the binary images are treated as gray-scale images in which image pixels take either the maximal or the minimal intensity values only.

Two observations motivate the use of complex wavelet coefficients. First, it has long been known that phase contains more structural information than magnitude in typical natural images [110]. Second, rigid translation of image structures leads to consistent phase shift. Based on these observations, it is preferable to design a metric that 1) separates the measurement of magnitude and phase distortions; 2) is more sensitive to phase than magnitude distortions; and 3) is insensitive to consistent *relative* phase distortions. In the complex wavelet transform domain (see [111] for an example), let $\mathbf{c}_x = \{c_{x,i} | i = 1, \dots, N\}$ and $\mathbf{c}_y = \{c_{y,i} | i = 1, \dots, N\}$ be two sets of coefficients extracted at the same spatial location (e.g., within a local window) in the same wavelet subbands of the two images being compared, respectively. The CW-SSIM index is defined as:

$$\tilde{S}(\mathbf{c}_x, \mathbf{c}_y) = \frac{2 |\sum_{i=1}^N c_{x,i} c_{y,i}^*| + K}{\sum_{i=1}^N |c_{x,i}|^2 + \sum_{i=1}^N |c_{y,i}|^2 + K}, \quad (5.4)$$

where c^* denotes the complex conjugate of c and K is a small positive constant.

The CW-SSIM index can be written as a product of two components:

$$\tilde{S}(\mathbf{c}_x, \mathbf{c}_y) = \frac{2 \sum_{i=1}^N |c_{x,i}| |c_{y,i}| + K}{\sum_{i=1}^N |c_{x,i}|^2 + \sum_{i=1}^N |c_{y,i}|^2 + K} \cdot \frac{2 |\sum_{i=1}^N c_{x,i} c_{y,i}^*| + K}{2 \sum_{i=1}^N |c_{x,i} c_{y,i}^*| + K}. \quad (5.5)$$

The first component is completely determined by the magnitudes of the coefficients and the maximum value 1 is achieved if and only $|c_{x,i}| = |c_{y,i}|$ for all i 's. This is a relatively insensitive measure. For example, scaling the magnitude of all coefficients by a factor of 10% only causes reduction of the CW-SSIM value from 1 to 0.9955 when $K = 0$ and the reduction is even less when $K > 0$. By contrast, the second component is fully determined by the consistency of phase changes between \mathbf{c}_x and \mathbf{c}_y . It achieves the maximum value 1 when the phase difference between $c_{x,i}$ and $c_{y,i}$ is a constant for all i 's. On the other hand, when the phase change between $c_{x,i}$ and $c_{y,i}$ is significantly different for different i 's, this measure gives a value far less than 1. In other words, the second component penalizes changes in image structures (inconsistent phase shift) and is robust to image translations (consistent phase shift). When the metric is applied *locally*, the robustness to translation also results in robustness to small rigid geometrical distortions such as rotation and scaling, because small geometrical distortions can be locally approximated with translations [98].

To implement the CW-SSIM index for the comparison of images, we first decompose the images using a complex version of a 2-scale, 4-orientation steerable pyramid decomposition [111], which is a type of redundant wavelet transform that avoids aliasing in subbands. The CW-SSIM indices are then computed locally using a sliding 7x7 window that moves across each wavelet subband. Finally, the resulting CW-SSIM index map is combined into a scalar similarity measure using a weighted summation. The weights are obtained

using a Gaussian profile with a standard deviation equaling a quarter of the image size.

5.3 Data Description and Experimental Design

Similarity metrics were compared on a set of simulated images and on tracings of linear objects on mammograms made by two experienced radiologists. To generate the simulated images, a binary image was created (Fig. 5.3). The simulated image represents an idealized model of a spiculated mass as seen on mammography. Variants of the test image were created by rotation, translation, and adjusting the edge thickness. The original image was compared to each of the rotated and translated images and similarity metrics were computed. These pairs of images model the case in which two readers trace linear structures on images, but their markings are off by a few pixels.

In the first experiment, the original image was rotated from 0 to 4 degrees in increments of 0.4 degrees. Translation was not applied to these images and the edge width was fixed at one pixel. Three images that have been rotated by 1, 2 and 4 degrees are shown in Figs. 5.3(a), 5.3(b), and 5.3(c) respectively.

In the second experiment, spatial translation was applied to the original image by shifting in the x or y directions by a few pixels. Let x_{shift} and y_{shift} denote the amount of spatial translation applied in the x and y directions. The Euclidean distance is $= \sqrt{(x_{shift}^2 + y_{shift}^2)}$. Ten spatial translations of 1 to 7 pixels (Euclidean distance) were applied to the original image. Rotation

was not applied to these images and the edge width was fixed at one pixel. Three images that have been translated by (3,0), (5,1) and (7,2) pixels in the horizontal and vertical directions are shown in Figs. 5.4(a), 5.4(b), and 5.4(c) respectively.

In the third experiment, the effect of the width on the similarity metrics was studied. For this experiment, the line segments in the Fig. 5.5(a) were dilated once and twice with a 3-by-3 window. The two dilated images are shown in Figs. 5.5(b) and 5.5(c) respectively. A rotation of 4 degrees was applied to the images and the images were not translated.

The second set of images for this study were obtained from the Digital Database for Screening Mammography (DDSM) [62], the largest publicly available dataset of digitized mammograms. The outlines of the abnormalities are available in the database as “chain codes”. From this “chain code”, an ROI containing a lesion can be extracted. Two radiologists (R1 and R2) marked linear structures of interest on the images (spicules). To compute an estimate of the intra-observer agreement, R1 repeated the process. The analysis was conducted on regions-of-interest (ROIs) using the ROI Manager plugin of NIH ImageJ. The radiologists traced all spiculations on a set of 12 spiculated masses.

5.4 Results

5.4.1 Results on Simulated Data

5.4.1.1 Effect of rotation

We analyzed the effect of small rotations on intensity-based, spatial-based, and CW-SSIM metrics. In the simulation, the original image is rotated from 0 to 4 degrees in increments of 0.4 degrees and the original image is compared to each of the rotated images. Since the intensity-based metrics (Table 5.2) are all closely related, some are equivalent in our simulation. For the simulated images ($b = c$) and the Dice, Kulczynski2, Braun-Blanquet, Simpson, McConnaughey, and Ochiai metrics are equivalent and simplify to the formula $a/(a + b)$. Thus, Figure 5.3(d) shows the effect of rotation on the Dice and Sokal and Sneath(2) metrics. The values of the Dice and Sokal and Sneath(2) metrics decrease drastically as the amount of rotation is increased (Figs. 5.3(d)). In this simulation, the value of both metrics is zero for any rotation greater than 1.6 degrees. For example, the reported similarity between the original image and copy that is rotated 2 degrees (Fig. 5.5(b)) is the same as that between the original image and a copy that is rotated 4 degrees (Fig. 5.5(c)). Yet, there is certainly a perceptible difference between these scenarios and one may desire a metric that can quantify that difference. In comparison, the PCM has a greater dynamic range than the Dice metric does over rotational distortions. PCM is 0.67 for 2 degrees of rotation and 0.13 for 4 degrees of rotation (Fig. 5.3(e)). Similarly, the dynamic range of CW-SSIM is larger than that of PCM. The CW-SSIM metric has a value of 0.81 for 2 degrees of

rotation and a value of 0.65 for 4 degrees of rotation (Fig. 5.3(f)). A wider dynamic range may be advantageous in some applications. For example, the perceptible difference between the two simulations in Figs. 5.5(b) and 5.5(c) is slight. However, the PCM is 81% lower for the latter over the former, whereas, in comparison the CW-SSIM metric decreases by only 20% as the rotation increases from 2 to 4 degrees. Finally, note that the response of the CW-SSIM metric is linear, whereas it is not linear for Dice and PCM for these simulation settings. Human observer studies would be needed to determine if a linear or non-linear response better matches human perception of similarity for a given segmentation task. However, in the absence of such observer studies, the fact that the CW-SSIM values are easier to interpret is intuitively appealing.

5.4.1.2 Effect of translation

We analyzed the effect of small translations on intensity-based, spatial-based, and CW-SSIM metrics. For the simulated data, ten spatial translations of 1 to 7 pixels (Euclidean distance) were applied to the original image. The original image was then compared to each of the translated images. For the same reasons as given above, only the Dice and Sokal and Sneath(2) metrics are plotted. Both the Dice metric and the Sokal and Sneath(2) metric decrease significantly as the amount of translation is increased (Figs. 5.4(d)). In this simulation, the value of both metrics is zero for any translation greater than 3 units (Euclidean distance). For example, the reported similarity between the original image and a copy that is translated by approximately 5 units (Fig.

5.4(b)) is the same as that between the original image and a copy that is translated by approximately 7 units (Fig. 5.5(c)). Thus, the intensity-based metrics can't quantify such a difference in similarity. As was the case for rotational distortions, PCM has a greater dynamic range than the Dice metric does over translational distortions. For example, PCM is 0.43 for the first case and 0.22 for second case (Fig. 5.4(e)). Again, the dynamic range of CW-SSIM is larger than that of PCM. The CW-SSIM metric has a value of 0.70 for the first case and a value of 0.56 for the second (Fig. 5.4(f)). There is only a small perceptible difference between the two simulations in Figs. 5.4(b) and 5.4(c), yet the PCM decreases by 49% for the second case. In comparison the CW-SSIM metric decreases by only 20%. Finally, note that as was the case for the rotation simulation, the response of the CW-SSIM metric is linear as a function of translation, whereas it is not linear for Dice and PCM for these simulation settings.

5.4.1.3 Effect of width of edges

We analyzed the effect of width of edges on intensity-based, spatial based, and CW-SSIM metrics. In this experiment, only the width of the edges is changed, with the rotation fixed at 4 degrees and no translation applied. For the same reasons as given above, only the Dice and Sokal and Sneath(2) metrics are plotted. Both intensity-based (Fig. 5.5(d)) and spatial-based (Fig. 5.5(e)) metrics are very sensitive to the width of edges whereas the CW-SSIM metric is more robust to the changes in the width of edges (Fig. 5.5(f)).

We also studied the effect of the width of edges in combination with rotations or translations. The performance of the metrics for three different edge widths and various rotations are depicted in Figs. 5.6(b), 5.6(d) and 5.6(f). Similarly, the performance of the metrics for three different edge widths and various translations are depicted in Figs. 5.7(b), 5.7(d) and 5.7(f). The behavior of the Dice and PCM metrics as a function of rotation or translation depends on the width used to represent the segmentations. For smaller values of the edge-width, the responses of both Dice and PCM are non-linear. By comparison, for larger values of the edge-width, the responses of both Dice and PCM metrics appear linear. Additionally, the CW-SSIM metric has a larger dynamic range than Dice and PCM for all simulation conditions.

5.4.2 Results on Mammography Data

The two radiologists (R1 and R2) measured the properties of 12 spiculated masses. To compute the intra-observer agreement, R1 measured these properties twice. Quantitative measures of both the intra- and inter-observer agreement were computed using the Dice, PCM, and CW-SSIM metrics.

Two sets of outlines of spiculations made by one radiologist (R1) on four spiculated masses are shown in Fig. 5.8. Each row in this figure shows the original image, the two set of outlines generated by R1, and the values of the three metrics. By visual inspection, it is apparent that there is substantial intra-observer agreement in the segmentations of some images (e.g., Fig. 5.8(a)). However, low values of the Dice metric are obtained (e.g., Dice = 0.29

for Fig. 5.8(a)); note that a Dice value of greater than 0.7 is considered to denote “excellent” agreement [103]. In contrast, the PCM and CW-SSIM values are much higher for segmentations of images displaying obvious intra-observer agreement (e.g., PCM = 0.36 and CW-SSIM = 0.54, for Fig. 5.8(a)).

On the other hand, the relative values of all three metrics are generally consistent with subjective assessments. For example, by visual inspection, we can see that the intra-observer agreement in the segmentations of the spiculated mass shown in Fig. 5.8(a) is more than that for the spiculated mass in Fig. 5.8(m). This difference is apparent from each of the metrics: Dice = 0.29 vs. 0.09, PCM = 0.36 vs. 0.20, and CW-SSIM = 0.54 vs 0.34. However, there are some images for which the intra-observer agreement in segmentation was clearly greater than for other images, yet only the CW-SSIM metric captured this difference. For example, by visual inspection, we can see that the intra-observer agreement in the segmentations of the spiculated mass shown in Fig. 5.8(e) is more than that for the spiculated mass in Fig. 5.8(m). This difference is not apparent from the Dice and PCM metrics, but is captured by the CW-SSIM metric: Dice = 0.07 vs. 0.09, PCM = 0.19 vs. 0.20, and CW-SSIM = 0.48 vs. 0.34.

Two sets of outlines of spiculations made by two radiologists (R1 and R2) on four spiculated masses are shown in Fig. 5.9. Each row in this figure shows the original image, the outlines generated by R1 and R2, and the values of the three metrics. Again, low values of Dice are obtained despite subjectively obvious agreement. For example, the inter-observer agreement for the

spiculated mass shown in Fig. 5.9(i) is only 0.16 by Dice. In comparison, PCM (0.26) and CW-SSIM (0.49) indicate more substantial agreement. As was the case for intra-observer agreement, all metrics generally captured the trends in the inter-observer agreement expected from visual inspection. For example, by visual inspection, we can see that the inter-observer agreement in the segmentations of the spiculated mass shown in Fig. 5.9(i) is more than that for the spiculated mass in Fig. 5.9(m). All three metrics reflect this observation: Dice = 0.16 vs. 0.10, PCM = 0.26 vs. 0.25, and CW-SSIM = 0.49 vs. 0.34.

Finally, one would expect that the intra-observer agreement should be higher than the inter-observer agreement. All of the metrics exhibit this trend for most of the segmentation pairs. The Dice and PCM values are greater for intra-observer agreement than for the corresponding inter-observer agreement for 8 out of the 12 mammography cases and the CW-SSIM values are greater for intra-observer vs. inter-observer for 9 out of the 12 of the images of spiculated masses.

5.5 Discussion

In this chapter, we have compared metrics for assessing the similarity of segmentations of linear structures. The metrics include intensity-based measures such as Dice and spatial-based approaches such as PCM. We have introduced the use of the CW-SSIM to quantify image similarity. Simulations were conducted to study the effect of small geometric distortions such as rotation and translation on these similarity metrics. Likewise, the effect of the

thickness of edges was studied. The similarity metrics were also compared for the task of quantifying intra- and inter-observer agreement of segmentations of linear structures on mammograms.

Dice and other intensity-based metrics are simple and intuitive. They can be implemented easily and have low computational overhead. Consequently, intensity-based metrics are widely used in medical imaging. However, as we have shown with a number of experiments in this study, intensity-based metrics have a small dynamic range for comparing segmentations of linear structures.

Spatial-based metrics such as CDM and PCM have advantages over the intensity-based metrics for evaluating the segmentations of linear structures. In this study, we have demonstrated that spatial-based metrics have a larger dynamic range than intensity-based metrics. However, spatial-based metrics require one to specify a search neighborhood and it is difficult to do so upfront. In this analysis, a search neighborhood of 2 pixels was used in order to be consistent with a prior study by the developers of PCM [96]. It is likely that a using larger search neighborhood would help improve the performance of CDM and PCM for evaluating similarity of segmentations of linear structures. However, increasing the search area would also increase the computational complexity. As CDM and PCM perform exhaustive searches for matching pixels of the boundaries, the main limitation of these metrics is that they are computationally intensive.

In our simulations, CW-SSIM had the largest dynamic range for quanti-

fying dissimilarity due to small geometric transformations such as rotation and translation. Likewise, while the dynamic range of CW-SSIM did not depend on the edge width, the dynamic ranges of Dice and PCM were very sensitive to the edge width. Moreover, the responses of Dice and PCM as a function of rotational and translational distortion was nonlinear for small edge widths. In comparison, the response of CW-SSIM was linear for all edge widths (Figs. 5.6 and 5.7).

The Dice, PCM, and CW-SSIM metrics were also applied to mammography images in which observers traced spicules on spiculated masses. Both the intra- and inter-observer agreement was quantified. One would expect that the intra-observer agreement should be higher than the inter-observer agreement. All of the metrics exhibit this trend for most of the segmentation pairs. However, note that as with the simulated data, the absolute values of the Dice metric for the intra- and inter-observer agreement are low relative to the reported cutoff of 0.7 for “excellent” agreement. The absolute values of the PCM and CW-SSIM metrics are larger, but it is not possible to make definite conclusions about the dynamic range of these metrics on real data.

Assessing the similarity of segmentations is a challenging problem. While it is easy to define a multitude of potential metrics, it is difficult to evaluate their performance since there is not a “ground truth” for the concept of similarity. Through careful simulations and demonstrations with real data, such as we have presented in this study, one can provide some insight into the relative strengths and weakness of different similarity metrics. However,

how to best interpret the values computed for a given similarity metric remains elusive. In contrast, evaluation scales have been developed for metrics for assessing similarity or agreement in other tasks. For example, there are established scales for interpreting the intra-class correlation coefficient (ICC), which is used to assess observer agreement in measurements studies (e.g., compare observers' measurements of mass size), and the Kappa statistic, which is used to assess observer agreement in rating studies (e.g., compare observers' descriptions of mass margin). To the best of our knowledge, Dice is unique among similarity metrics for comparing segmentations in that a cut-off (0.7) for "excellent" similarity has been proposed (by Zijdenbos *et al.* [103]). The justification of the 0.7 cut-off for Dice is based on an appeal to its relationship with Kappa. However, as Zijdenbos *et al.* acknowledge, interpreting the absolute value of Dice is still difficult, even given a defined cut-off for excellent agreement. Thus, future work is needed to increase the interpretability of similarity metrics for comparing image segmentations.

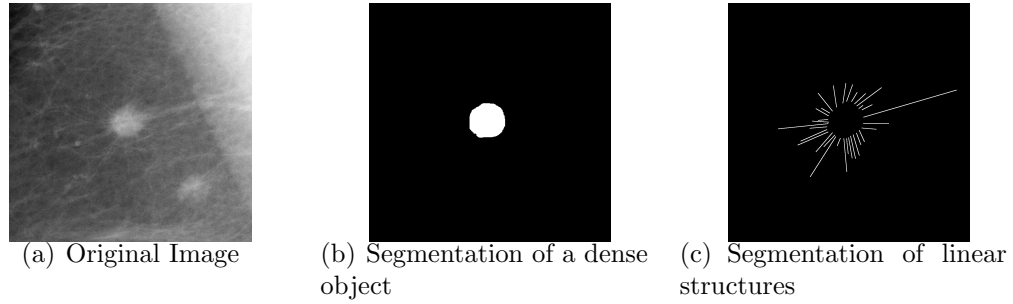


Figure 5.1: This figure shows examples of the two extremes of segmentations. In Fig 5.1(b) the central mass region (dense object) of a spiculated mass is segmented. In Fig. 5.1(c) the spicules (linear structures) are outlined. The segmentations are represented by setting the pixels on or inside the boundary to one and zero otherwise.

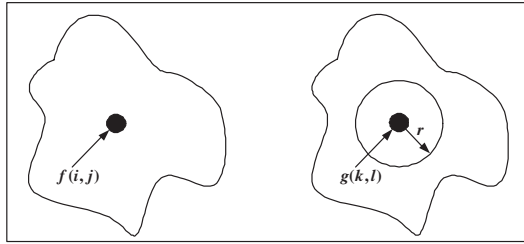


Figure 5.2: A schematic explanation of the PCM algorithm. Let f and g represent the two segmentation images that are to be compared. Let $f(i, j)$ represent a particular pixel in f and $g(k, l)$ represent the corresponding pixel in g . For every pixel $f(i, j)$, the PCM algorithm tries to find its match within a neighborhood of radius r of the corresponding pixel $g(k, l)$.

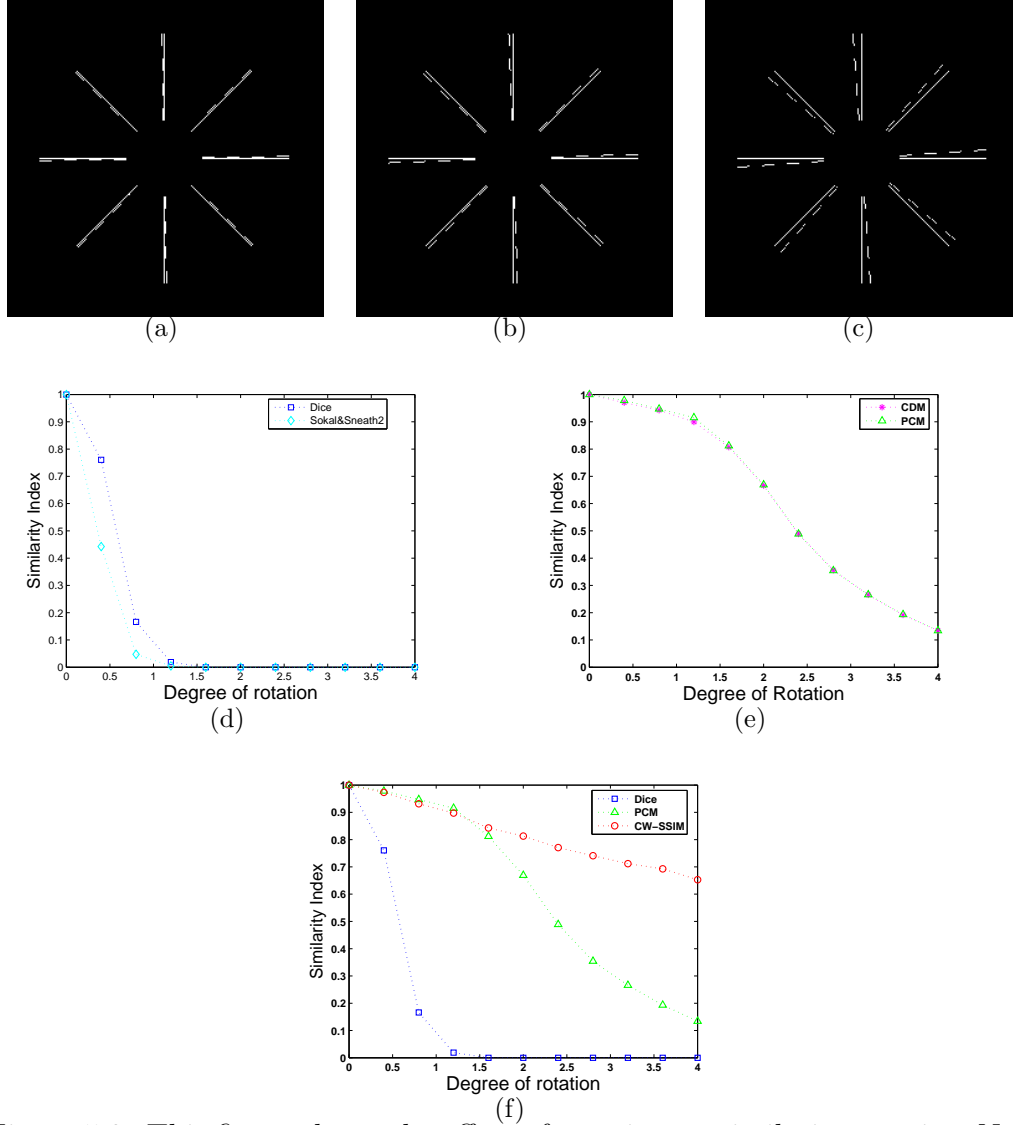


Figure 5.3: This figure shows the effect of rotation on similarity metrics. Note that the edges widths are fixed and no translation is applied. In Figs. 5.3(a), 5.3(b), and 5.3(c) the original image is displayed with solid lines. The locations of the lines of the image that was obtained after applying a rotation is shown in these images with dotted lines. In Figs. 5.3(a), 5.3(b), and 5.3(c) rotations of 1, 2, and 4 degrees were applied respectively. Figure 5.3(d) shows the effect of rotation on the intensity-based metrics whereas Fig. 5.3(e) compares the performance of the spatial-based metrics for different rotations. The performance of the Dice (intensity-based), PCM (spatial-based), and CW-SSIM are compared in Fig. 5.3(f).

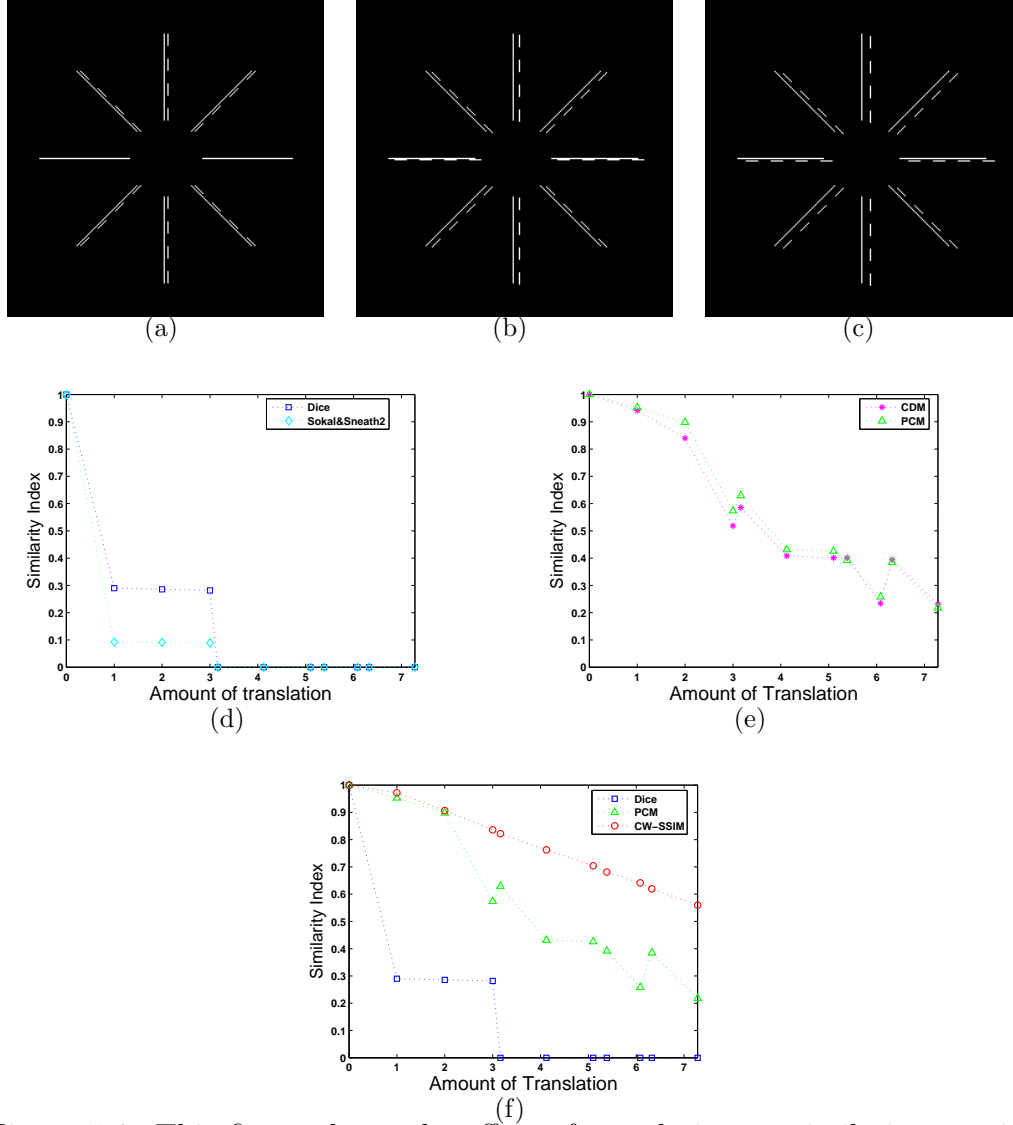


Figure 5.4: This figure shows the effect of translation on similarity metrics. Note that the edges widths are fixed and no rotation is applied. In Figs. 5.4(a), 5.4(b), and 5.4(c) the original image is displayed with solid lines. The locations of the lines of the image that was obtained after applying a translation is shown in these images with dotted lines. In Figs. 5.4(a), 5.4(b), and 5.4(c) translations of 3.00, 5.10, and 7.28 units (Euclidean distance) were applied respectively. Figure 5.4(d) shows the effect of translation on the intensity-based metrics whereas Fig. 5.4(e) compares the performance of the spatial-based metrics for different translations. The performance of Dice (intensity-based), PCM (spatial-based), and CW-SSIM are compared in Fig. 5.4(f).

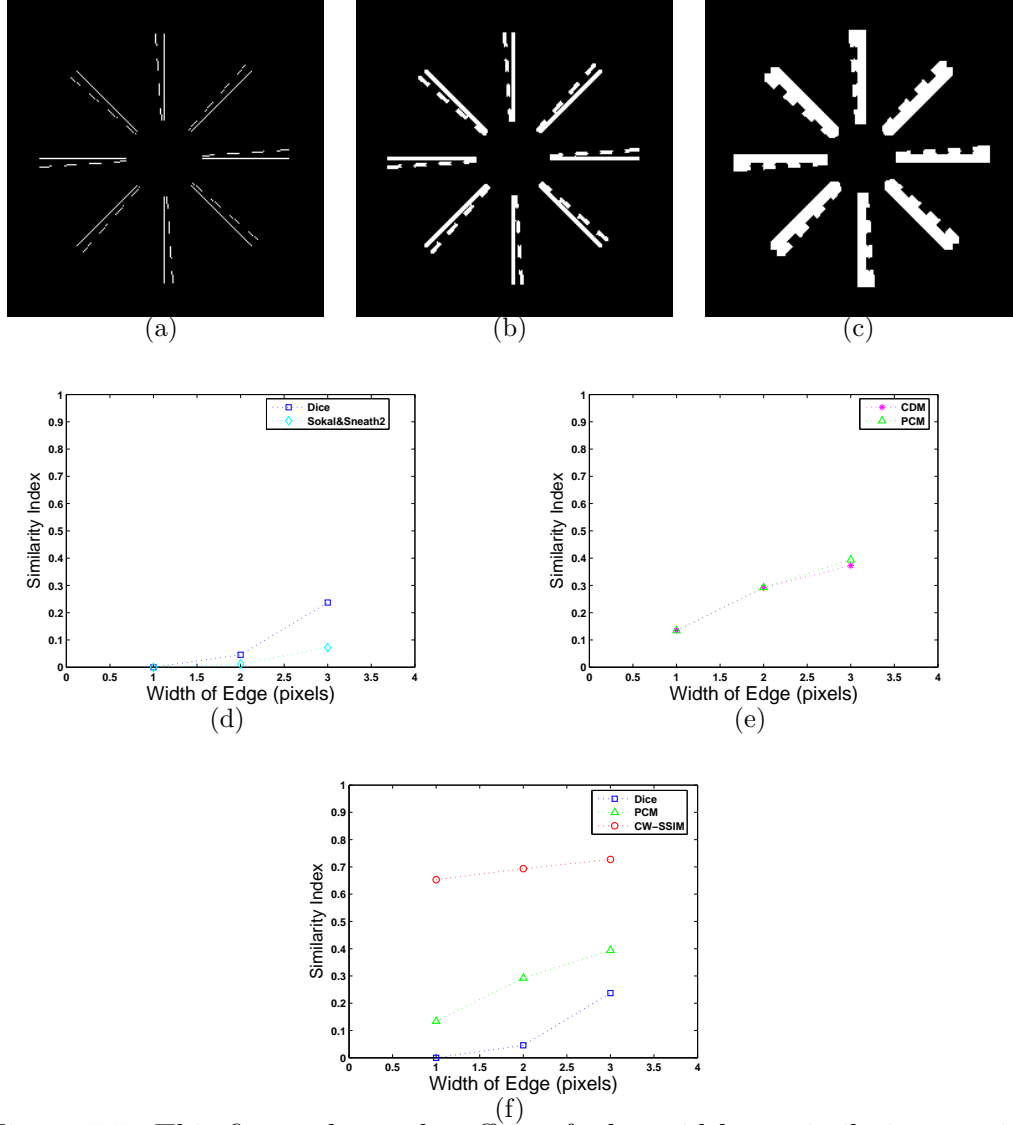
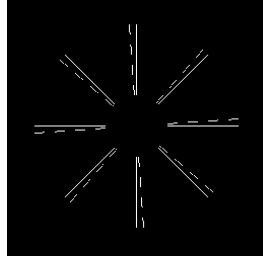
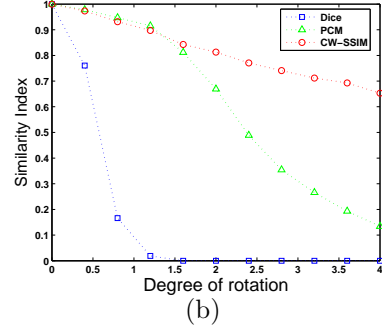


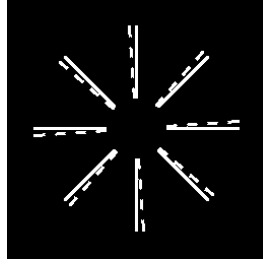
Figure 5.5: This figure shows the effect of edge width on similarity metrics. Note that a fixed rotation of 4 degrees was applied to the original image and no translation was applied in this simulation. In Fig. 5.5(a) the edge width was one pixel, whereas the simulated image was dilated by factors of 1 and 2 in Figs. 5.5(b) and 5.5(c) respectively. Figure 5.5(d) shows the effect of edge width on the intensity-based metrics whereas Fig. 5.5(e) compares the performance of the spatial-based metrics for different edge widths. The performance of Dice (intensity-based), PCM (spatial-based), and CW-SSIM are compared in Fig. 5.5(f).



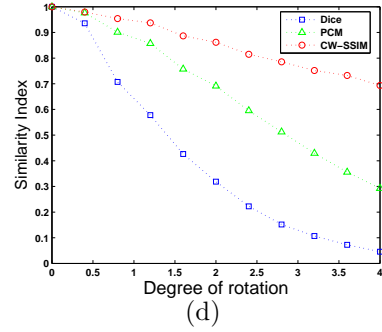
(a)



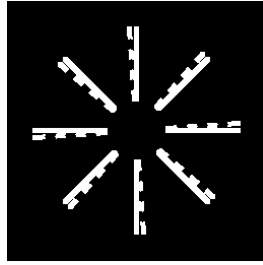
(b)



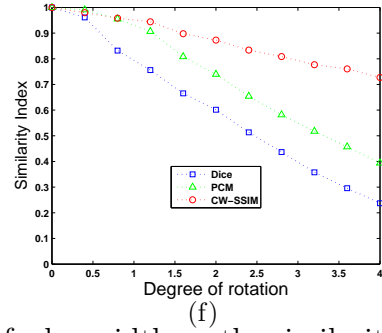
(c)



(d)

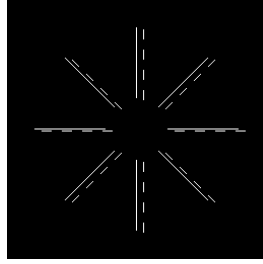


(e)

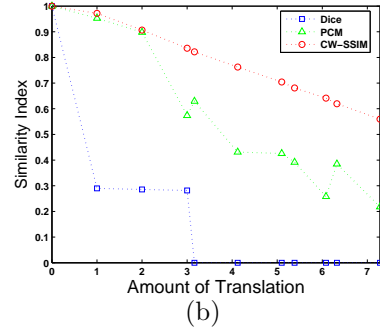


(f)

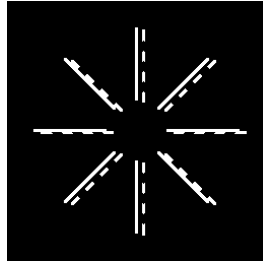
Figure 5.6: This figure shows the effect of edge width on the similarity metrics as a function of rotation distortion. The original image was rotated from 0 to 4 degrees in increments of 0.4 degrees. The locations of the lines of the image that was obtained after applying a rotation are shown with dotted lines (for example, in Fig. 5.6(a), the original image is rotated by 4 degrees). The original image was compared to each of the rotated images and for each pair of images the similarity metrics were computed. To study the effect of the edge width, the image was dilated and the metrics were recalculated. Figures 5.6(c) and 5.6(e) show the simulated images after dilation by factors of 1 and 2 respectively and Figs. 5.6(d) and 5.6(f) show the similarity metrics as a function of rotation.



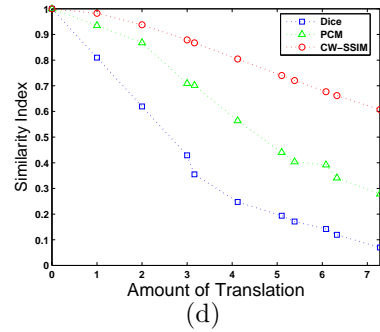
(a)



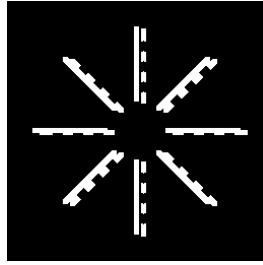
(b)



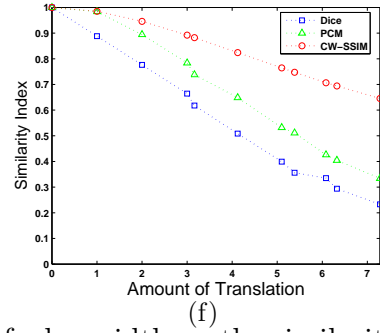
(c)



(d)



(e)



(f)

Figure 5.7: This figure shows the effect of edge width on the similarity metrics as a function of translation distortion. The original image was translated in the x and y directions by different amounts. The locations of the lines of the image that was obtained after applying spatial translations are shown with dotted lines (for example, in Fig. 5.7(a) a translation of 7 pixels to the right and 2 pixels to the bottom is applied). The original image was compared to each of the translated images and for each pair of images the similarity metrics were computed. To study the effect of the edge width on the metrics, the image was dilated and the metrics recalculated. Figures 5.7(c) and 5.7(e) show the simulated images after dilation by factors of 1 and 2 respectively and Figs. 5.7(d) and 5.7(f) show the similarity metrics as a function of translation.

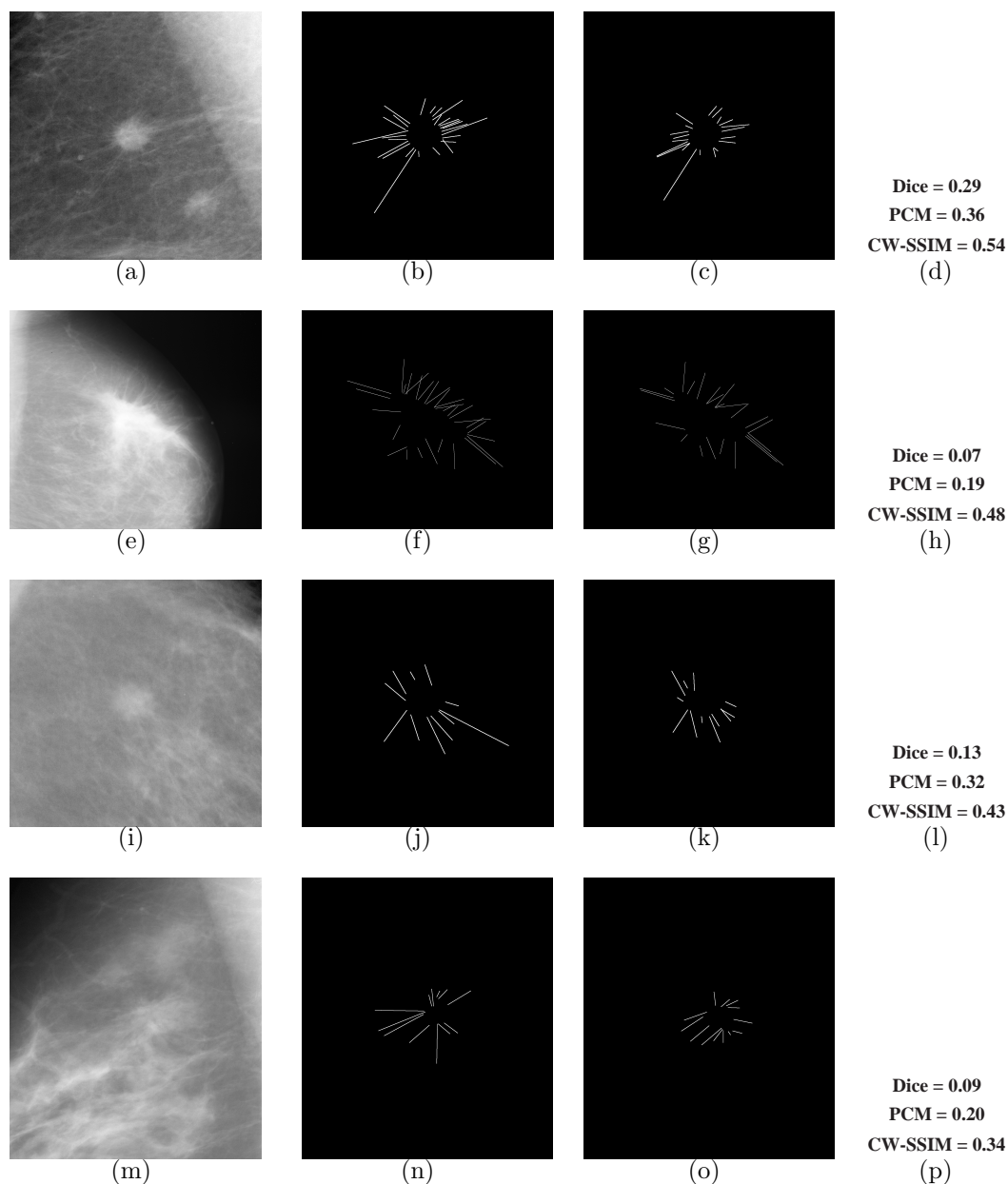


Figure 5.8: Intra-observer agreement: This figure shows two sets of outlines of spicules created by a radiologist (R1) for four spiculated masses. The first column shows the original images and the first and second sets of outlines made by R1 are displayed in columns two and three. Column four shows the Dice, PCM, and CW-SSIM metrics for each pair of segmentations.

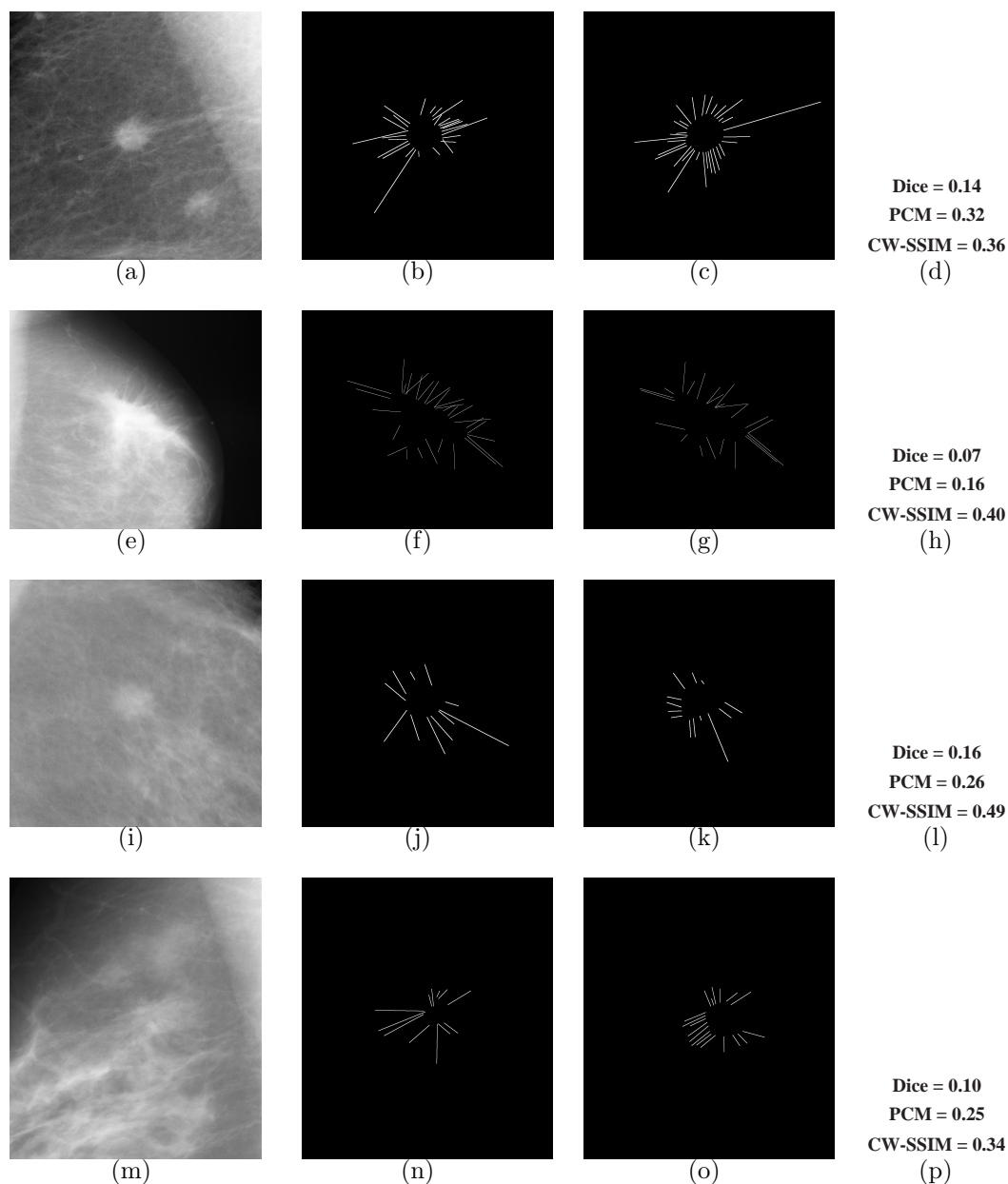


Figure 5.9: Inter-observer agreement: This figure shows the outlines of spicules created by two radiologists (R1 and R2) for four spiculated masses. The first column shows the original images and the second and third columns show the outlines made by R1 and R2 respectively. Column four shows the Dice, PCM, and CW-SSIM metrics for each pair of segmentations.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this dissertation a new *evidence-based* detection algorithm for the detection of spiculated lesions on mammography was developed. By *evidence-based* we mean that we used the statistics of the physical characteristics of spiculated lesions to select the parameters of the detection algorithm.

For this task, a measurement study was conducted in which, for the first time, measurements on the physical properties of spiculated lesions were collected and used to guide the design of the detection algorithm. To the best of our knowledge, this is the first time the parameters of a detection algorithm have been selected from scientific data rather than from anecdotal impressions or empirically selected to optimize performance of a particular dataset. In this dissertation, measurements taken from this study were used to demonstrate the engineering principles that we have proposed. It was also shown (with various statistical techniques) that the properties of these lesions can be reliably measured and the first database of the physical properties of spiculated lesions was created.

In the development of the detection algorithm, we have made several

novel contributions. We have invented a new class of filter banks to detect the spatial location where these spicules converge. We refer to these as Spiculation Filter banks (SFBs). In our approach the SFBs are designed to explicitly match the expected structure of the spiculations (*viz.*, the SFs look like idealized spiculated lesions) while they also have desirable optimal spatio-spectral localization properties. Beyond this, the SFBs are not only designed to match the spiculated lesions to be encountered, but they are configured, based on the scientific data from the measurement studies, to have sizes, frequencies, and extents derived from actual physical measurements.

A new Radon-domain technique for enhancing spicules in mammograms was invented. The enhancement is achieved by computing the Radon transform of the image and filtering in the Radon domain. The parameters of the filter are chosen based on the width of spicules obtained from the measurement studies.

6.2 Future Work

6.2.1 Relations to Spiculation filter-banks to wavelet theory

Through appropriate selection of the Spiculation filter-banks (SFB) parameters, the SFB can be made to implement a specific type of wavelet transform. Thus, it is appropriate to think of the filters as Spiculation Wavelets and which can be used to compose Spiculation Wavelet Transforms. Many wavelet-based approaches for analyzing mammograms have been reported [22, 112–114] and the use of wavelets has proven to be a significant advance, owing to the

good approximation properties of appropriate wavelet functions. However, the wavelet basis functions used previously were standard wavelets, such as the Daubechies 9/7 [115, 116] which were not matched to the structures to be detected. Because of this, prior approaches could only use the wavelet outputs/coefficients as generic features for training a standard classifier, with the user hoping that the results would be adequate. To date, the results have not been adequate for spiculated lesions. We believe that taking a direct modeling approach will greatly increase the utility of wavelet-based algorithms and allow simple classification algorithms to be effective. In our approach, the wavelet functions are designed to explicitly match the expected structure of the spiculations (*viz.*, the SF look like idealized spiculated lesions) while they also have desirable optimal spatio-spectral localization properties. Further, the SFB are not only designed to match the spiculated lesions to be encountered, but they are configured to have sizes, frequencies, and extents derived from actual physical measurements. Since SFs comprise a radically different wavelet basis with properties and appearances that have never been explored previously or exploited before, this will offer an important avenue of study that may contribute significantly to wavelet theory.

6.2.2 Local Radon Transform

The new spicule enhancement algorithm presented here uses the Radon transform which can be viewed as a “global” transform. (Since image intensities are summed along lines and not along line segments.) Recently, a local-

ized Radon transform has been developed [117–119]. We plan to investigate whether localized Radon transforms help to improve the accuracy of the overall detection method by more selective, local feature enhancement.

6.2.3 Measuring properties of other linear structures in mammograms

In this dissertation, the physical properties of spiculated lesions were measured and used. A number of other linear structures such as blood vessels, ducts and fibrous tissue and connective tissue can mimic the appearance of spicules. Some researchers have tried to distinguish spicules from such linear structures. For example, Zwiggelaar *et al.* proposed a technique in which various linear structures detected on mammograms were classified into categories such as spicules, vessels, fibrous tissue, etc [120]. However, currently, there are no data on the properties of linear structures such as vessels, ducts, and connective tissue that are commonly seen on mammograms. Measuring these properties would help to select the parameters of classification algorithms in a principled manner.

6.2.4 Collection of measurement data on spiculated lesions

In this work, measurements were made by two radiologists on a set of 21 spiculated masses and 15 architectural distortions. I believe it would be beneficial to collect measurements on a much larger set of spiculated masses and architectural distortions and from multiple radiologists. Once we have these measurements, these should be made available to the research community.

A comprehensive set of statistics of the physical parameters of spiculated lesions could be used not only for improving detection algorithms but also for more localized tasks such as diagnosis and reduction of false positives. I strongly believe that the systematic study and quantification of the physical parameters of these lesions would not only be useful for our developmental efforts but also for those of future researchers in computer-aided detection and diagnosis.

Appendices

Appendix A

List of images used for measurement study

This appendix provides the details of the images used for the measurement studies. The names of the spiculated masses used for the measurement studies are provided in Table A.1 and the names of the architectural distortions used are shown in Table A.2. All of these images were obtained from the DDSM database [62].

B_3135_1.RIGHT_MLO.LJPEG	C_0016_1.RIGHT_MLO.LJPEG
B_3071_1.LEFT_MLO.LJPEG	B_3405_1.RIGHT_MLO.LJPEG
C_0095_1.RIGHT_MLO.LJPEG	C_0158_1.LEFT_MLO.LJPEG
C_0139_1.RIGHT_MLO.LJPEG	C_0175_1.LEFT_MLO.LJPEG
C_0339_1.RIGHT_MLO.LJPEG	C_0180_1.LEFT_MLO.LJPEG
C_0081_1.LEFT_MLO.LJPEG	C_0069_1.RIGHT_MLO.LJPEG
C_0103_1.LEFT_MLO.LJPEG	C_0181_1.RIGHT_MLO.LJPEG
C_0188_1.LEFT_MLO.LJPEG	C_0096_1.RIGHT_MLO.LJPEG
C_0110_1.RIGHT_MLO.LJPEG	C_0165_1.LEFT_MLO.LJPEG
B_3016_1.RIGHT_MLO.LJPEG	C_0004_1.RIGHT_MLO.LJPEG
C_0358_1.LEFT_MLO.LJPEG	

Table A.1: This table lists the set of spiculated masses used in the measurement studies. Each radiologist measured the properties of these spiculated masses.

C_0020_1.RIGHT_MLO.LJPEG	B_3425_1.LEFT_MLO.LJPEG
C_0481_1.LEFT_MLO.LJPEG	C_0032_1.RIGHT_MLO.LJPEG
B_3139_1.LEFT_MLO.LJPEG	C_0137_1.LEFT_MLO.LJPEG
C_0101_1.RIGHT_MLO.LJPEG	B_3475_1.LEFT_MLO.LJPEG
C_0070_1.LEFT_MLO.LJPEG	B_3514_1.LEFT_MLO.LJPEG
C_0183_1.RIGHT_MLO.LJPEG	C_0026_1.LEFT_MLO.LJPEG
C_0073_1.RIGHT_MLO.LJPEG	C_0171_1.LEFT_MLO.LJPEG
B_3093_1.LEFT_MLO.LJPEG	A_1674_1.LEFT_MLO.LJPEG

Table A.2: This table lists the set of architectural distortions used in the measurement studies. Each radiologist measured the properties of these architectural distortions.

Appendix B

List of images used for the detection algorithm

This appendix provides the details of the images used to test the performance of the detection algorithm. Table B.1 lists the names of the spiculated masses and Table B.2 lists the names of the architectural distortions used. All of these images were obtained from the DDSM database [62]. Images of spiculated masses from the MIAS database ([79]) were also used. Table B.3 lists the names of these image.

B_3012_1.RIGHT_MLO.LJPEG	B_3041_1.LEFT_MLO.LJPEG
B_3046_1.LEFT_MLO.LJPEG	B_3057_1.RIGHT_MLO.LJPEG
B_3065_1.RIGHT_MLO.LJPEG	B_3084_1.RIGHT_MLO.LJPEG
B_3398_1.RIGHT_MLO.LJPEG	B_3411_1.RIGHT_MLO.LJPEG
B_3492_1.LEFT_MLO.LJPEG	C_0003_1.RIGHT_MLO.LJPEG
C_0004_1.RIGHT_MLO.LJPEG	C_0009_1.RIGHT_MLO.LJPEG
C_0015_1.RIGHT_MLO.LJPEG	C_0018_1.LEFT_MLO.LJPEG
C_0028_1.LEFT_MLO.LJPEG	C_0034_1.RIGHT_MLO.LJPEG
C_0047_1.LEFT_MLO.LJPEG	C_0050_1.LEFT_MLO.LJPEG
C_0059_1.LEFT_MLO.LJPEG	C_0064_1.RIGHT_MLO.LJPEG
C_0074_1.LEFT_MLO.LJPEG	C_0075_1.LEFT_MLO.LJPEG
C_0079_1.LEFT_MLO.LJPEG	C_0085_1.RIGHT_MLO.LJPEG
C_0096_1.RIGHT_MLO.LJPEG	C_0100_1.RIGHT_MLO.LJPEG
C_0103_1.LEFT_MLO.LJPEG	C_0110_1.RIGHT_MLO.LJPEG
C_0112_1.LEFT_MLO.LJPEG	C_0116_1.LEFT_MLO.LJPEG
C_0121_1.RIGHT_MLO.LJPEG	C_0141_1.LEFT_MLO.LJPEG
C_0158_1.LEFT_MLO.LJPEG	C_0165_1.LEFT_MLO.LJPEG
C_0166_1.RIGHT_MLO.LJPEG	C_0170_1.LEFT_MLO.LJPEG
C_0180_1.LEFT_MLO.LJPEG	C_0181_1.RIGHT_MLO.LJPEG
C_0187_1.LEFT_MLO.LJPEG	C_0192_1.RIGHT_MLO.LJPEG
C_0194_1.RIGHT_MLO.LJPEG	C_0196_1.LEFT_MLO.LJPEG
C_0225_1.RIGHT_MLO.LJPEG	C_0226_1.RIGHT_MLO.LJPEG
C_0339_1.RIGHT_MLO.LJPEG	C_0340_1.RIGHT_MLO.LJPEG
C_0358_1.LEFT_MLO.LJPEG	C_0424_1.RIGHT_MLO.LJPEG
C_0432_1.LEFT_MLO.LJPEG	C_0457_1.RIGHT_MLO.LJPEG

Table B.1: This table lists the set of spiculated masses used in the detection studies. The detection algorithm is applied to each of these images and results are reported as FROC curves.

A_1127_1.RIGHT_MLO.LJPEG	A_1140_1.LEFT_MLO.LJPEG
A_1203_1.LEFT_MLO.LJPEG	A_1217_1.LEFT_MLO.LJPEG
A_1222_1.RIGHT_MLO.LJPEG	A_1260_1.LEFT_MLO.LJPEG
A_1282_1.LEFT_MLO.LJPEG	A_1388_1.RIGHT_MLO.LJPEG
A_1401_1.RIGHT_MLO.LJPEG	A_1410_1.LEFT_MLO.LJPEG
A_1436_1.LEFT_MLO.LJPEG	A_1467_1.RIGHT_MLO.LJPEG
A_1478_1.LEFT_MLO.LJPEG	A_1504_1.RIGHT_MLO.LJPEG
A_1505_1.LEFT_MLO.LJPEG	A_1517_1.RIGHT_MLO.LJPEG
A_1577_1.RIGHT_MLO.LJPEG	A_1587_1.LEFT_MLO.LJPEG
A_1592_1.RIGHT_MLO.LJPEG	A_1628_1.LEFT_MLO.LJPEG
A_1720_1.LEFT_MLO.LJPEG	A_1720_1.RIGHT_MLO.LJPEG
A_1777_1.LEFT_MLO.LJPEG	A_1804_1.LEFT_MLO.LJPEG
A_1835_1.RIGHT_MLO.LJPEG	A_1847_1.RIGHT_MLO.LJPEG
A_1906_1.RIGHT_MLO.LJPEG	B_3123_1.LEFT_MLO.LJPEG
B_3440_1.RIGHT_MLO.LJPEG	B_3477_1.LEFT_MLO.LJPEG
B_3495_1.LEFT_MLO.LJPEG	B_3501_1.RIGHT_MLO.LJPEG
C_0006_1.LEFT_MLO.LJPEG	C_0142_1.LEFT_MLO.LJPEG
C_0230_1.RIGHT_MLO.LJPEG	C_0283_1.LEFT_MLO.LJPEG
C_0418_1.RIGHT_MLO.LJPEG	D_4059_1.RIGHT_MLO.LJPEG
D_4077_1.LEFT_MLO.LJPEG	D_4102_1.RIGHT_MLO.LJPEG
D_4137_1.LEFT_MLO.LJPEG	D_4142_1.RIGHT_MLO.LJPEG
D_4143_1.LEFT_MLO.LJPEG	D_4150_1.RIGHT_MLO.LJPEG
D_4160_1.LEFT_MLO.LJPEG	D_4173_1.LEFT_MLO.LJPEG

Table B.2: This table lists the set of architectural distortions used in the detection studies. The detection algorithm is applied to each of these images and results are reported as FROC curves.

mdb145.pgm	mdb148.pgm
mdb175.pgm	mdb178.pgm
mdb179.pgm	mdb181.pgm
mdb184.pgm	mdb186.pgm
mdb188.pgm	mdb190.pgm
mdb191.pgm	mdb193.pgm
mdb195.pgm	mdb198.pgm
mdb199.pgm	mdb202.pgm
mdb204.pgm	mdb206.pgm
mdb207.pgm	

Table B.3: This table lists the set of spiculated masses (from the MIAS database) used in the detection studies. The detection algorithm is applied to each of these images and results are reported as FROC curves.

Bibliography

- [1] American Cancer Society. Cancer facts and figures 2006. Technical report, American Cancer Society, 2006.
- [2] E. J. Feuer, L. M. Wun, C. C. Boring, W. D. Flanders, M. J. Timmel, and T. Tong. The lifetime risk of developing breast cancer. *J Natl Cancer Inst*, 85(11):892–897, Jun 1993.
- [3] L. M. Wun, R. M. Merrill, and E. J. Feuer. Estimating lifetime and age-conditional probabilities of developing cancer. *Lifetime Data Anal*, 4(2):169–186, 1998.
- [4] American College of Radiology. *ACR BI-RADS - Mammography, Ultrasound & Magnetic Resonance Imaging*. American College of Radiology, Reston, VA, fourth edition, 2003.
- [5] S. G. Orel, N. Kay, C. Reynolds, and D. C. Sullivan. BI-RADS categorization as a predictor of malignancy. *Radiology*, 211(3):845–850, Jun 1999.
- [6] Carol H Lee. Screening mammography: proven benefit, continued controversy. *Radiol Clin North Am*, 40(3):395–407, May 2002.
- [7] K. Kerlikowske, P. A. Carney, B. Geller, M. T. Mandelson, S. H. Taplin, K. Malvin, V. Ernster, N. Urban, G. Cutter, R. Rosenberg, and

- R. Ballard-Barbash. Performance of screening mammography among women with and without a first-degree relative with breast cancer. *Ann Intern Med*, 133(11):855–863, Dec 2000.
- [8] Thomas M Kolb, Jacob Lichy, and Jeffrey H Newhouse. Comparison of the performance of screening mammography, physical examination, and breast us and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. *Radiology*, 225(1):165–175, Oct 2002.
- [9] R. E. Bird, T. W. Wallace, and B. C. Yankaskas. Analysis of cancers missed at screening mammography. *Radiology*, 184(3):613–617, Sep 1992.
- [10] Maryellen L Giger. Computer-aided diagnosis in radiology. *Acad Radiol*, 9(1):1–3, Jan 2002.
- [11] D. B. Kopans. The positive predictive value of mammography. *AJR Am J Roentgenol*, 158(3):521–526, Mar 1992.
- [12] M. L. Giger, N. Karssemeijer, and S. G. Armato. Computer-aided diagnosis in medical imaging. *IEEE Trans Med Imaging*, 20(12):1205–1208, Dec 2001.
- [13] K. Doi, H. MacMahon, S. Katsuragawa, R. M. Nishikawa, and Y. Jiang. Computer-aided diagnosis in radiology: potential and pitfalls. *Eur J Radiol*, 31(2):97–109, Aug 1999.

- [14] T. W. Freer and M. J. Ulissey. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology*, 220(3):781–786, Sep 2001.
- [15] C. J. Vyborny, T. Doi, K. F. O’Shaughnessy, H. M. Romsdahl, A. C. Schneider, and A. A. Stein. Breast cancer: importance of spiculation in computer-aided detection. *Radiology*, 215(3):703–707, Jun 2000.
- [16] C. J. Vyborny, M. L. Giger, and R. M. Nishikawa. Computer-aided detection and diagnosis of breast cancer. *Radiol Clin North Am*, 38(4):725–740, Jul 2000.
- [17] David Gur, Jules H Sumkin, Howard E Rockette, Marie Ganott, Christiane Hakim, Lara Hardesty, William R Poller, Ratan Shah, and Luisa Wallace. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. *J Natl Cancer Inst*, 96(3):185–190, February 2004.
- [18] Dev P Chakraborty and Kevin S Berbaum. Observer studies involving detection and localization: modeling, analysis, and validation. *Med Phys*, 31(8):2313–2330, Aug 2004.
- [19] H.D. Li, M. Kallergi, L.P. Clarke, V.K. Jain, and R.A. Clark. Markov random field for tumor detection in digital mammography. *Medical Imaging, IEEE Transactions on*, 14(3):565–576, 1995.

- [20] W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper. Computer-aided mammographic screening for spiculated lesions. *Radiology*, 191(2):331–337, May 1994.
- [21] N. Karssemeijer and G.M. te Brake. Detection of stellate distortions in mammograms. *Medical Imaging, IEEE Transactions on*, 15(5):611–619, 1996.
- [22] Sheng Liu, C.F. Babbs, and E.J. Delp. Multiresolution detection of spiculated lesions in digital mammograms. *Image Processing, IEEE Transactions on*, 10(6):874–884, 2001.
- [23] T. Matsubara, H. Fujita, T. Endo, K. Horita, M. Ikeda, C. Kido, and T. Ishigaki. Development of mass detection algorithm based on adaptive thresholding technique in digital mammograms. In *Digital Mammography*, pages 391–296, June 1996.
- [24] H. Li, Y. Wang, K. J. Liu, S. C. Lo, and M. T. Freedman. Computerized radiographic mass detection—part I: Lesion site selection by morphological enhancement and contextual segmentation. *IEEE Trans Med Imaging*, 20(4):289–301, Apr 2001.
- [25] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley-InterScience, San Diego, Second edition, November 2000.

- [26] H. Kobatake, M. Murakami, H. Takeo, and S. Nawano. Computerized detection of malignant tumors on digital mammograms. *IEEE Trans Med Imaging*, 18(5):369–378, May 1999.
- [27] Nicholas Petrick, Heang-Ping Chan, Berkman Sahiner, and Datong Wei. An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection. *IEEE Transactions on Medical Imaging*, 15(1):59–67, February 1996.
- [28] N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie. Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. *Med Phys*, 26(8):1642–1654, Aug 1999.
- [29] W. E. Polakowski, D. A. Cournoyer, S. K. Rogers, M. P. DeSimio, D. W. Ruck, J. W. Hoffmeister, and R. A. Raines. Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency. *IEEE Trans Med Imaging*, 16(6):811–819, Dec 1997.
- [30] D.B. Kopans. *Breast Imaging*. Lippincott Williams, 2nd edition, January 1998.
- [31] D. Brzakovic, X. M. Luo, and P. Brzakovic. An approach to automated detection of tumors in mammograms. *IEEE Transactions on Medical Imaging*, 9(3):233–241, September 1990.

- [32] W Qian, L Li, L Clarke, R A Clark, and J Thomas. Digital mammography: comparison of adaptive and nonadaptive cad methods for mass detection. *Acad Radiol*, 6(8):471–480, August 1999.
- [33] S.-M. Lai, X. Li, and W.F. Biscof. On techniques for detecting circumscribed masses in mammograms. *Medical Imaging, IEEE Transactions on*, 8(4):377–386, Dec. 1989.
- [34] Bennett R. Groshong and W. Philip Kegelmeyer. Evaluation of a hough transform method for circumscribed lesion detection. *Digital Mammography*, pages 361–366, 1996.
- [35] M. Zhang, M. L. Giger, C. J. Vyborny, and K. Doi. Mammographic texture analysis for the detection of spiculated lesions. In K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, editors, *Digital Mammography*, pages 347–350. Elsevier, 1996.
- [36] B. Sahiner, Heang-Ping Chan, N. Petrick, Datong Wei, M.A. Helvie, D.D. Adler, and M.M. Goodsitt. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *Medical Imaging, IEEE Transactions on*, 15(5):598–610, Oct. 1996.
- [37] D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt. Classification of mass and normal breast tissue on digital mammograms: multiresolution texture analysis. *Med Phys*, 22(9):1501–1513, Sep 1995.

- [38] D. Wei, H. P. Chan, N. Petrick, B. Sahiner, M. A. Helvie, D. D. Adler, and M. M. Goodsitt. False-positive reduction technique for detection of masses on digital mammograms: global and local multiresolution texture analysis. *Med Phys*, 24(6):903–914, Jun 1997.
- [39] G. M. te Brake, N. Karssemeijer, and J. H. Hendriks. An automatic method to discriminate malignant masses from normal tissue in digital mammograms. *Phys Med Biol*, 45(10):2843–2857, Oct 2000.
- [40] M.A. Kupinski and M.L. Giger. Investigation of regularized neural networks for the computerized detection of mass lesions in digital mammograms. In *Engineering in Medicine and Biology society, 1997. Proceedings of the 19th Annual International Conference of the IEEE*, volume 3, pages 1336–1339, 1997.
- [41] Georgia D Tourassi, Rene Vargas-Voracek, David M Catarious, and Carey E Floyd. Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information. *Med Phys*, 30(8):2123–2130, Aug 2003.
- [42] Alan H Baydush, David M Catarious, Craig K Abbey, and Carey E Floyd. Computer aided detection of masses in mammography using subregion hotelling observers. *Med Phys*, 30(7):1781–1787, Jul 2003.
- [43] T. Hara, T. Matsubara, T. Inazawa, and H. et al. Makita. Automated detection method for architectural distortion based on distribution as-

- essment of mammary gland on mammogram. In *8th International Workshop on Computer-Aided Diagnosis*, 2006.
- [44] R.M. Rangayyan and F. Ayres. Detection of architectural distortion in mammograms using a shape-constrained phase portrait model. In *8th International Workshop on Computer-Aided Diagnosis*, 2006.
 - [45] A.S. Elmaghraby, N.H. Eltonsy, and G.D. Tourassi. Investigating performance of a morphology-based cad scheme in detecting architectural distortion in screening mammograms. 2006.
 - [46] Summary of Safety and Effectiveness Data: R2 Technologies. P970058. U.S. Food and Drug Administration, 1998.
 - [47] Summary of Safety and Effectiveness Data: ISSI. P010038. U.S. Food and Drug Administration, 2002.
 - [48] Summary of Safety and Effectiveness Data: CADx Medical Systems. P010038. U.S. Food and Drug Administration, 2002.
 - [49] B. Zheng, M. A. Gannot, and C. A. Britton et.al. Soft-Copy Mammographic Readings with Different Computer-assisted Detection Cuing Environments: Preliminary Findings. *Radiology*, 221:633–640, 2001.
 - [50] Jay A Baker, Eric L Rosen, Joseph Y Lo, Edgardo I Gimenez, Ruth Walsh, and Mary Scott Soo. Computer-aided detection (CAD) in screening mammography: sensitivity of commercial CAD systems for

- detecting architectural distortion. *AJR Am J Roentgenol*, 181(4):1083–1088, Oct 2003.
- [51] Carl J. D’Orsi. Computer-aided Detection: There Is No Free Lunch. *Radiology*, 221:585–586, 2001.
- [52] K. Woods and K. Bowyer. A general view of detection algorithms. *Digital Mammography*, pages 385–390, June 1996.
- [53] M. P. Sampat, G. J. Whitman, M. K. Markey, and A. C. Bovik. Evidence based detection of spiculated lesions and architectural distortions. In *SPIE Medical Imaging, Image Processing*, volume 5747, pages 26–37, Feb 2005.
- [54] W. A. Berg, C. Campassi, P. Langenberg, and M. J. Sexton. Breast imaging reporting and data system: inter- and intraobserver variability in feature analysis and final assessment. *AJR Am J Roentgenol*, 174(6):1769–1777, Jun 2000.
- [55] M. J. Vos, B. M J Uitdehaag, F. Barkhof, J. J. Heimans, H. C. Baayen, W. Boogerd, J. A. Castelijns, P. H M Elkhuisen, and T. J. Postma. Interobserver variability in the radiological assessment of response to chemotherapy in glioma. *Neurology*, 60(5):826–830, Mar 2003.
- [56] Jeremy J Erasmus, Gregory W Gladish, Lyle Broemeling, Bradley S Sabloff, Mylene T Truong, Roy S Herbst, and Reginald F Munden. Interobserver and intraobserver variability in measurement of non-small-

cell carcinoma lung lesions: implications for assessment of tumor response. *J Clin Oncol*, 21(13):2574–2582, Jul 2003.

- [57] Naama R Bogot, Ella A Kazerooni, Aine M Kelly, Leslie E Quint, Benoit Desjardins, and Bin Nan. Interobserver and intraobserver variability in the assessment of pulmonary nodule size on CT using film and computer display methods. *Acad Radiol*, 12(8):948–956, Aug 2005.
- [58] L. Valentin and I. Bergelin. Intra- and interobserver reproducibility of ultrasound measurements of cervical length and width in the second and third trimesters of pregnancy. *Ultrasound Obstet Gynecol*, 20(3):256–262, Sep 2002.
- [59] Marianne de Vries, Patrick J de Koning, Michiel W de Haan, Alphons G Kessels, Patricia J Nelemans, Robbert J Nijenhuis, R. Nils Planken, G. Boudewijn C Vasbinder, Jos M A van Engelshoven, Rob J van der Geest, and Tim Leiner. Accuracy of semiautomated analysis of 3D contrast-enhanced magnetic resonance angiography for detection and quantification of aortoiliac stenoses. *Invest Radiol*, 40(8):495–503, Aug 2005.
- [60] Stephan G Wetzel, Soonmee Cha, Glyn Johnson, Peter Lee, Meng Law, David L Kasow, Sean D Pierce, and Xiaonan Xue. Relative cerebral blood volume measurements in intracranial mass lesions: interobserver and intraobserver reproducibility study. *Radiology*, 224(3):797–803, Sep 2002.

- [61] K. D. Hopper, C. J. Kasales, M. A. Van Slyke, T. A. Schwartz, T. R. TenHave, and J. A. Jozefiak. Analysis of interobserver and intraobserver variability in CT tumor measurements. *AJR Am J Roentgenol*, 167(4):851–854, Oct 1996.
- [62] M. Heath, K. Bowyer, and D. Kopans. Current status of the digital database for screening mammography. In *Digital Mammography*, pages 457–460. Kluwer Academic Publishers, 1998.
- [63] Ronir Raggio Luiz and Moyses Szklo. More than one statistical strategy to assess agreement of quantitative measurements may usefully be reported. *J Clin Epidemiol*, 58(3):215–216, Mar 2005.
- [64] K. O. McGraw and S. P. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46, 1996.
- [65] J. M. Bland and D. G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1(8476):307–310, Feb 1986.
- [66] J. M. Bland and D. G. Altman. Applying the right statistics: analyses of measurement studies. *Ultrasound Obstet Gynecol*, 22(1):85–93, Jul 2003.
- [67] S. Wellek. *Testing statistical hypothesis of equivalence*. CRC Press LLC, Boca Raton, FL, 2003.
- [68] B. Rosner. *Fundamentals of Biostatistics*. Duxbury Press, 2005.

- [69] A. C. Bovik. *Handbook of Image and Video Engineering*. Academic Press, 2nd edition, 2005.
- [70] Athanasios Papoulis. *The Fourier Integral and Its Applications*. McGraw-Hill, 1962.
- [71] W. A. Gotz and H. J. Druckmuller. A fast digital radon transform- an efficient means for evaluating the hough transform. *Pattern Recognition*, 28(12):1985–1992, 1995.
- [72] Azriel Rosenfeld and Avinash C. Kak. *Digital Picture Processing*, volume 1. Academic Press, second edition, 1982.
- [73] A.C. Kak and M. Slaney. *Principles of computerized tomographic imaging*. IEEE Press, 1988.
- [74] M. P. Sampat and A. C. Bovik. Detection of spiculated lesions in mammograms. In *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, volume 1, pages 810–813 Vol.1, 2003.
- [75] A.C. Bovik, M. Clark, and W.S. Geisler. Multichannel texture analysis using localized spatial filters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(1):55–73, 1990.
- [76] A.C. Bovik. Analysis of multichannel narrow-band filters for image texture segmentation. *Signal Processing, IEEE Transactions on*, 39(9):2025–2043, 1991.

- [77] A.C. Bovik, N. Gopal, T. Emmoth, and A. Restrepo. Localized measurement of emergent image frequencies by gabor wavelets. *Information Theory, IEEE Transactions on*, 38(2):691–712, 1992.
- [78] M. Clark, A. Bovik, and W. Geisler. Texture segmentation using a class of narrowband filters. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.*, volume 12, pages 571–574, 1987.
- [79] J Suckling, J Parker, D Dance, and S Astley et al. The mammographic images analysis society digital mammogram database. In *Excerpta Medica. International Congress Series*, pages 375–378, 1994.
- [80] R. Zwiggelaar, T. C. Parr, J. E. Schumm, I. W. Hutt, C. J. Taylor, S. M. Astley, and C. R. Boggis. Model-based detection of spiculated lesions in mammograms. *Med Image Anal*, 3(1):39–62, Mar 1999.
- [81] A. Averbuch, R. Coifman, D. Donoho, and W. Israel. Fast slant stack: A notion of radon transform for data in a cartesian grid which is rapidly computible, algebraically exact, geometrically faithful and invertible. Technical report, Stanford University, 2001.
- [82] M. P. Sampat, M. K. Markey, and A. C. Bovik. *Handbook of Image and Video Processing*, chapter Computer-Aided Detection and Diagnosis in Mammography, pages 1195–1217. Academic Press, 2nd edition, 2005.

- [83] J. Udupa, V. Leblanc, H. Schmidt, C. Imielinska, P. Saha, G. Grevera, Y. Zhuge, L. Currie, P. Molholt, and Y. Jin. Methodology for evaluating image-segmentation algorithms. In *Medical Imaging, Proceedings of SPIE.*, 2002.
- [84] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, Jul. 1945.
- [85] P. Jaccard. The distribution of flora in the alpine zone. *New Phytologis*, 11:37–50, 1912.
- [86] G.G Simpson. Notes on the measurement of faunal resemblance. *Am. J. Sci.*, 258:300–311, 1960.
- [87] A. Ochiai. Zoogeographic studies on the soleoid fishes found in japan and its neighbouring regions. *Bull. Jpn. Soc. Sci. Fish*, 22:526–530, 1957.
- [88] J. Braun-Blanquet. *Plant Sociology: The Study of Plant Communities*. McGraw Hill, 1932.
- [89] R.R. Sokal and P.A. Sneath. *Principles of Numerical Taxonomy*. W. H. Freeman, 1963.
- [90] R.R. Sokal and C.D. Michener. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.*, 38:1409–1438, 1958.

- [91] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.
- [92] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [93] J. Jomier, V. LeDigarcher, and S. Aylward. Comparison of vessel segmentations using staple. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 523–530, 2005.
- [94] William Pratt. *Digital Image Processing*. John Wiley, 2001.
- [95] K. Bowyer, C. Kranenburg, and S. Dougherty. Edge detector evaluation using empirical roc curves. *Computer Vision and Image Understanding*, 84(1):77–103, 2001.
- [96] M.S. Prieto and A.R. Allen. A similarity metric for edge images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(10):1265–1273, 2003.
- [97] Liang Shen, R.M. Rangayyan, and J.E.L. Desautels. Application of shape analysis to mammographic calcifications. *Medical Imaging, IEEE Transactions on*, 13(2):263–274, 1994.

- [98] Z. Wang and E.P. Simoncelli. Translation insensitive image similarity in complex wavelet domain. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 573–576, 2005.
- [99] Guang R. Shi. Multivariate data analysis in palaeoecology and palaeobiogeography—a review. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 105(3-4):199–234, November 1993.
- [100] J. M. Duarte, J. B. Santos, and L. C. Melo. Comparison of similarity coefficients based on rapd markers in the common bean. *Genet. Mol. Biol.*, 22:427–432, 1999.
- [101] D. A. Jackson, K. M. Somers, and H. H. Harvey. Similarity coefficients: Measures of co-occurrence and association or simply measures of occurrence? *American Naturalist*, 133(3):436–453, Mar. 1989.
- [102] K. Zou, S. Warfield, A. Bharatha, et al. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol*, 11(2):178–189, February 2004.
- [103] A.P. Zijdenbos, B.M. Dawant, R.A. Margolin, and A.C. Palmer. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Transactions on Medical Imaging*, 13(4):716–724, 1994.
- [104] A. Bharatha, M. Hirose, N. Hata, S. Warfield, and others. Evaluation of three-dimensional finite element-based deformable registration of pre-

- and intraoperative prostate imaging. *Med. Phys.*, 28(12):2551–2560, December 2001.
- [105] S. Kulczynski. Zespoly rslin w pieninach. *Bull. Int. Acad. Pol. Sci. Lettres*, 2:57–203, 1928.
- [106] Z. Hubalek. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biological Reviews*, 57(4):669–689, 1982.
- [107] P.F. Russel and T.R. Rao. On habitat and association of species of anopheline larvae in south-eastern madras. *J. Malaria Inst. India*, 3:153–178, 1940.
- [108] U.G. Yule. On the association of attributes in statistics. *Philosophical Transactions of the Royal Society of London*, 194:257319, 1900.
- [109] D.J. Rogers and T.T. Tanimoto. A computer program for classifying plants. *Science*, 132:1115–1118, 1960.
- [110] A.V. Oppenheim and J.S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.
- [111] J. Portilla and E.P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–71, 2000.

- [112] R.J.; Ferrari, R.M.; Rangayyan, J.E.L.; Desautels, and A.F.; Frere. Analysis of asymmetry in mammograms via directional filtering with gabor wavelets. *IEEE Transactions on Medical Imaging*, 20(9):953–964, 2001.
- [113] A.F. Laine, S. Schuler, Jian Fan, and W. Huda. Mammographic feature enhancement by multiscale analysis. *Medical Imaging, IEEE Transactions on*, 13(4):725–740, 1994.
- [114] L. Li, F. Mao, W. Qian, and L.P. Clarke. Wavelet transform for directional feature extraction in medical imaging. In *Image Processing, 1997. Proceedings., International Conference on*, volume 3, pages 500–503, 1997.
- [115] Ingrid Daubechies. *Ten Lectures on Wavelets*. Soc for Industrial & Applied Math, 1992.
- [116] R. Raghuveer and A. Bopardikar. *Wavelet Transforms: Introduction to theory and applications*. Prentice Hall, 1998.
- [117] A.C. Copeland, G. Ravichandran, and M.M. Trivedi. Localized radon transform-based detection of ship wakes in sarimages. *Geoscience and Remote Sensing, IEEE Transactions on*, 33(1):35–45, 1995.
- [118] A.L. Warrick and P.A. Delaney. Detection of linear features using a localized radon transform. In *Signals, Systems and Computers, 1996*.

1996 Conference Record of the Thirtieth Asilomar Conference on, pages 1245–1249 vol.2, 1996.

- [119] A.L. Warrick and P.A. Delaney. Detection of linear features using a localized radon transform with a wavelet filter. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 4, pages 2769–2772 vol.4, 1997.
- [120] R. Zwiggelaar, S.M. Astley, C.R.M. Boggis, and C.J. Taylor. Linear structures in mammographic images: detection and classification. *Medical Imaging, IEEE Transactions on*, 23(9):1077–1086, Sept. 2004.

Vita

Mehul Pravin Sampat was born in Mumbai, India on 24 Nov 1978. He graduated from the University of Mumbai, Mumbai in May 2000 with a B.E. degree in Biomedical Engineering. He joined the Dept. of Biomedical Engineering at the University of Texas at Austin (UT-Austin) in August 2000 and obtained a masters degree in Biomedical Engineering in May 2002. Since Spring 2001 he has been a Graduate Research Assistant at the Laboratory for Image and Video Engineering under the supervision of Dr. Alan C. Bovik. He joined the Ph.D. program in Biomedical Engineering in Sept. 2003 and has been supervised by Dr. Alan C. Bovik and Dr. Mia K. Markey. He is also a member of the Biomedical Informatics Lab at UT-Austin since Sept. 2003.

Permanent address: 304-A, Rokadia Apts., Rokadia Lane, Borivali-West-Mumbai-400092

This dissertation was typeset with \LaTeX^\dagger by the author.

[†] \LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's \TeX Program.